Social and Personality Psychology Compass 10/10 (2016), 535–549, 10.1111/spc3.12267

Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How

Jin X. Goh[1]*, Judith A. Hall[1] and Robert Rosenthal[2]

[1] *Northeastern University*
[2] *University of California, Riverside*

Abstract

We outline the need to, and provide a guide on how to, conduct a meta-analysis on one's own studies within a manuscript. Although conducting a "mini meta" within one's manuscript has been argued for in the past, this practice is still relatively rare and adoption is slow. We believe two deterrents are responsible. First, researchers may not think that it is legitimate to do a meta-analysis on a small number of studies. Second, researchers may think a meta-analysis is too complicated to do without expert knowledge or guidance. We dispel these two misconceptions by (1) offering arguments on why researchers should be encouraged to do mini metas, (2) citing previous articles that have conducted such analyses to good effect, and (3) providing a user-friendly guide on calculating some meta-analytic procedures that are appropriate when there are only a few studies. We provide formulas for calculating effect sizes and converting effect sizes from one metric to another (e.g., from Cohen's *d* to *r*), as well as annotated Excel spreadsheets and a step-by-step guide on how to conduct a simple meta-analysis. A series of related studies can be strengthened and better understood if accompanied by a mini meta-analysis.

How many studies does one need to conduct a meta-analysis? A perusal of meta-analyses published in 2016 in *Psychological Bulletin* showed that the number of studies included ranged from 63 (Williams & Tiedens, 2016) to 243 (Pool, Brosch, Delplanque, & Sander, 2016). Meta-analysis can certainly be associated with a daunting number of studies, and its capacity to analyze a large number of studies is one of its virtues, especially if the goal is to uncover moderator variables. However, while having a large number of studies can serve a researcher well, it is not a prerequisite. Statistically speaking, only two values are needed to calculate an arithmetic mean. In the same vein, only two studies are needed to conduct a meta-analysis (more precisely, only two effect sizes or two *p*-values are needed). In fact, some meta-analytic procedures involving only two effect sizes have been in use for decades but are not typically thought of as such – for example, the *Z*-test for comparing two independent correlation coefficients (McNemar, 1962).

The fact that meta-analysis can be done on a small number of studies opens up new opportunities for researchers to understand and strengthen their conclusions based on their own studies. It is easy for researchers, reviewers, editors, and readers to be confused and even to disagree over what the "bottom line" conclusion should be across several studies. What if three studies all produce the same directional trend, but one is $p = .05$, one is $p = .06$, and one is $p = .20$? What if the studies have different sample sizes? Or one goes in an unpredicted direction? Meta-analysis redirects attention towards effect sizes and away from individual studies' *p*-values, which have very limited comparative value. Even with a small number of studies, meta-analytic procedures allow one to summarize them, which not only clarifies the picture but leverages the statistical power provided by a meta-analysis. Especially when the effect size that is being detected is small, individual studies may be underpowered and therefore unpersuasive according to standard inferential methods but persuasive when analyzed together. Furthermore, even with a small number of studies one can compare them (or subgroups of them) to each other. The goal of the present article is to persuade researchers to perform an internal meta-analysis – which we call a *mini meta* – whenever they have a series of conceptually comparable studies.

Conducting a meta-analysis on only a handful of studies is particularly useful given that the inclusion of multiple studies within one manuscript is a growing trend in psychological science. Regardless of how many studies were conducted, a researcher can succinctly summarize the findings across studies with a meta-analysis. Although a meta-analytic summary of the findings obtained may be more persuasive than considering each effect individually, this practice is still relatively rare among social and personality psychologists. For this reason, the current article outlines the need to and a guide on how to conduct a mini meta of one's own studies.

We are certainly not the first to argue for the benefits of a meta-analysis that is internal to one's own research. Maner (2014) recommended that reviewers and editors embrace inconsistency across multiple studies by encouraging authors to conduct meta-analysis on the author's own results. Braver, Thoemmes, and Rosenthal (2014) suggested that a continuous cumulation of effects through a meta-analysis can provide stronger evidence for the replicability of a research question than claiming evidence from a single "failed" replication (however "failed" might be defined). Cumming (2014) argued that conducting meta-analyses within manuscripts can increase precision of estimates (i.e., narrower confidence intervals).

We believe two deterrents are responsible for some researchers' unwillingness to conduct meta-analyses on their own studies. First, researchers may not think that it is worthwhile or legitimate to do a meta-analysis on a small number of studies. Second, researchers may think they lack the expertise to do a meta-analysis. To dispel these two misconceptions, we first provide arguments for conducting a mini meta within a manuscript (and show that it is actually appropriate to do so). We then provide a primer on how to do a small meta-analysis so that researchers will have the basic skills to conduct one. With only a few studies, the process is remarkably simple and requires no special software. One can use software, of course, such as the commercially available Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2005). All of the procedures we mention are standard in most meta-analysis textbooks (e.g., Rosenthal, 1991; Lipsey & Wilson, 2001). We do not attempt to cover every possible effect size metric, type of research design, or type of statistical analysis that one might face in the studies to be meta-analyzed, but the examples and resources we provide should be sufficient to get researchers started. Numerous resources are available for researchers who seek deeper understanding or whose research requires methods not covered in the present article; for more information in one book, see Cooper, Hedges, and Valentine (2009).

Some Arguments on Why

John, Loewenstein, and Prelec (2012) surveyed psychologists regarding their research practices and found that in this growing trend to incorporate multiple studies within a manuscript, one popular strategy in light of null findings is to suppress them in the (metaphorical and/or literal) file-drawer and only report the significant findings. The field's obsession with an arbitrary alpha level has generated a wave of critiques over the years (Cumming, 2014), and by limiting a paper to only significant findings, researchers will inevitably contribute to an unhealthy culture of questionable research practices such as *p*-hacking (Simonsohn, Nelson, & Simmons, 2014), file-drawer effects (Rosenthal, 1979), and more. The overall effect size obtained through a meta-analysis places more weight on the reliability and the replicability of findings than on individual effects that may or may not meet the .05 convention (Braver et al., 2014). Therefore, conducting a mini meta can provide greater transparency as researchers can include their "null" findings in the manuscript and still provide justification for their overall result.[1] If the practice of doing meta-analysis were to become a norm, then questionable research practices might be greatly reduced.

Furthermore, given that adequate power is often required for submission to top-tier journals (e.g., Lindsay, 2015; Vazire, 2016) and for grant funding, a mini meta uniquely offers researchers the necessary ingredients to calculate power. When an article contains a meta-analytic summary of the studies contained in it, researchers can use the given overall effect size to estimate the sample size needed for a high-power replication or a grant proposal on a similar question.

Another advantage of doing a mini meta is finding effects that are only detectable cumulatively rather than in a single study. Because certain phenomena can have small effects that are difficult to detect (as is often the case in personality and social psychology), one can accumulate the available findings that are often non-significant to provide a convincing argument that such an effect is indeed real via a mini meta. Rule and colleagues (2015) conducted a meta-analysis of 23 unpublished studies done in their respective laboratories that examined the relationship between anti-gay prejudice and accuracy in judging sexual orientation. While there were negative relationships in most of the studies, only three of these correlations were significant. Through a meta-analysis, however, a significant and homogeneous effect was found to confirm the researchers' hypothesis that stronger anti-gay prejudice was associated with worsened accuracy in detecting sexual orientation.

A meta-analysis can also be used to support arguments for null findings. Suppose one suspects there are no non-negligible differences between two groups or a relationship does not exist. Traditional inferential statistics cannot support the null hypothesis without unrealistically large sample sizes, but if the cumulative evidence overwhelmingly suggests a negligible effect size or a non-significant combined probability, this can be convincing. Hall and colleagues (2009) conducted 11 experiments to examine the effect of motivation to be accurate on interpersonal accuracy tests. Neither the individual studies nor the meta-analysis yielded a significant effect. Furthermore the overall effect size was near zero and homogeneous. Therefore, a meta-analysis can lend support to a near-null hypothesis when individual studies cannot.

When multiple studies using comparable measures are performed to validate a new test or scale, they can be amalgamated using meta-analysis. This makes not just for a more robust and convincing conclusion, but also for economical organization because results can be presented together (for examples, see Hall, Goh, Schmid Mast, & Hagedorn, 2016, or Goh, Schlegel, Tignor, & Hall, 2016). In these articles, each study's methodology was separately described, followed by an integrated presentation of the results. This organization is not only concise but lightens readers' cognitive burden because they can see all of the comparable results together instead of reading them serially and trying to hold them in memory one study to the next. Similar examples of employing a mini meta to summarize test validation results can be found in Rosenthal, Hall, DiMatteo, Rogers, and Archer (1979), Hall et al. (2015), and Carter, Hall, Carney, and Rosip (2006).

Mini metas have also been used to support counterintuitive findings. Lai and colleagues (in press) meta-analyzed two studies with nine different interventions designed to reduce implicit racial bias. Their mini metas showed that while these interventions were effective immediately in reducing prejudice, none of the interventions was actually effective after a delay (of several hours to several days). Meta-analytic effect sizes of the nine interventions at Time 2 (after a delay) showed that all interventions centered closely to $d = .00$. While it has been established that implicit prejudice is malleable, Lai and colleagues showed that the malleability has a time limit.

Mini metas can be and have been conducted on an even smaller scale with two studies alone. Hugenberg and Bodenhausen (2004) meta-analyzed two studies to strengthen their argument that implicit prejudice was associated with racial categorization when anger, but not happiness, was displayed. Williams and DeSteno (2008) found a significant relationship between pride and self-esteem in a mini meta of two studies. Lamarche and Murray (2014) did the same with two experiments that examined how people's self-esteem influenced attention in response to relationship threats, as did Case, Conlon, and Maner (2015) for two experiments on powerlessness and motivation to seek social affiliation. Lim and DeSteno (2016) found a non-significant relationship between perspective taking and compassion in one study but a significant relationship in another, and combining the two studies supported their argument that perspective taking was correlated with compassion. While it is encouraging to see that some researchers have included mini metas in their articles, the practice of meta-analyzing one's own studies is still rare.

A Primer on How

A second deterrent is the possibility that not all authors possess the expertise to conduct a meta-analysis. Therefore, we present a step-by-step guide below. This primer is not meant to be exhaustive, but it does provide researchers with the tools (and hopefully confidence) to conduct a simple meta-analysis. In our experience, people are quite surprised to learn how easy it is to calculate meta-analytic summaries and comparisons. All of these procedures are easy to do, even with a hand calculator, and involve only a few easily obtained ingredients. An annotated Excel spreadsheet is provided.[2] It will become obvious that this primer does not cover all contingencies, and that many potential subtleties and complexities in research design and statistics are not discussed. Fortunately, many online and published resources are available.

Background

As mentioned before, only two effect sizes are needed to calculate a meta-analysis, which is essentially an arithmetic mean (which may or may not be weighted according to study characteristics, such as sample size) when combining studies, or the difference between effect sizes when comparing them. An effect size (ES) measures the magnitude of an effect. There are numerous ES metrics available to researchers, but the two most common in psychology are Cohen's *d* (and other members of that family, such as Hedges' *g*; see Borenstein, Hedges, Higgins, & Rothstein, 2009), typically used for expressing the difference between two means, and the Pearson correlation (*r*) for expressing the linear relationship between two ordered variables. In the present article, we give illustrations only for *d* and *r*. These metrics are readily converted one to the other, which is not surprising considering that differences between groups and relationships between variables are merely different ways of talking about the same thing. Cohen (1992), Rosnow and Rosenthal (2009), and Lakens (2013) provide primers on ES, including procedures for deriving them from the results provided in the original studies.

Here we introduce both fixed effects and random effects approaches (Borenstein et al., 2009; Borenstein, Hedges, Higgins, & Rothstein, 2010). Fixed effects are usually used when the author believes there is one true population ES, which is most likely when studies are similar methodologically. Fixed effects are also more often used when the author wishes to make a statement about the studies on hand, without generalizing to new studies (with new populations, research instruments, etc.). The fixed approach assumes that all of the variance in ES across studies is due to sampling variation. Sampling variation can, of course, yield wildly discrepant ES (see demonstration in Schmidt, 1996). Given that studies within a manuscript are often very similar in their methods and the goal is to summarize those studies, some researchers might opt for the fixed approach. A second approach is the random effects approach, which is described in Step 7 below. This approach is generally very conservative if only a few studies are available, but it may be useful to compute because it affords greater generalizability. So should you use fixed effects or random effects? That depends on your goal for the meta-analysis. Both methods succeed in summarizing the findings. But if the goal is to generalize beyond your findings and you don't assume there is only one true ES, random effects may be a better approach. We recommend reporting both.[3]

In this primer, we use two examples to help clarify the steps: Jacob and Sarah. (A third researcher, Gordon, will be introduced later to discuss common ambiguities faced by meta-analysts.) The fabricated data of their studies are presented in Tables 1 and 2. You can follow the steps by analyzing their data.

Table 1. Jacob's data on gender differences of social network size.

	t	***df***	***p***	**Cohen's *d***	***r***
Study 1 ($N = 30$)	1.69	28	.102	.64	.30
Study 2 ($N = 50$)	2.01	48	.050	.58	.28
Study 3 ($N = 100$)	1.98	98	.051	.40	.20
$M\ r_z$					.24
$M\ r$					.24
Combined Z					3.25***

Correlations in the last column were calculated from t values using Formula 3. $M\ r_z$ = weighted mean correlation (Fisher's z transformed). $M\ r$ = weighted mean correlation (converted from r_z to r). In all analyses, men were coded as 0 and women as 1. Positive Cohen's d and positive correlation coefficients indicate that women have larger network size relative to men.
***$p < .001$, two-tailed.

Table 2. Sarah's data on correlations between age and Big Five traits.

	Agreeable	**Conscientious**	**Extraverted**	**Open**	**Neurotic**
Study 1 ($N = 80$)	.05	.17	.17	.04	.32
Study 2 ($N = 200$)	.40	.38	.24	−.19	.50
$M\ r_z$	.32	.34	.22	−.13	.49
$M\ r$	.31	.33	.22	−.13	.45
Combined Z	4.32***	4.88***	3.48***	-1.65^{+}	7.02***

$M\ r_z$ = weighted mean correlation (Fisher's z transformed). $M\ r$ = weighted mean correlation (converted from r_z to r).
$^{+}p < .10$, two-tailed.
***$p < .001$, two-tailed.

Step 1: Decide on your Research Question

In the present discussion, we assume that the studies to be combined address the same conceptual, directional hypothesis with the same or conceptually similar measures. A study often includes more than two variables, and one could potentially conduct several mini metas within a manuscript as well, as exemplified in Sarah's case below.

Jacob. Jacob wants to examine gender differences in size of social network (i.e., number of friends), and he did three studies. Gender differences were calculated using independent *t*-tests. In each ES, Jacob must decide what a positive or negative value would signify, and he must be consistent in applying his decision. He can arbitrarily code gender as 0 = male and 1 = female, and any ES with positive value would therefore indicate that women have a larger social network than men, and any ES with negative value would indicate that men have a larger social network than women. See Table 1.

Sarah. Sarah wants to examine how Big Five traits (extraversion, openness, neuroticism, agreeableness, and conscientiousness) correlate with age in two studies. Each study would have five ES, one for each trait. She would therefore do five separate meta-analyses to combine her two studies, one meta-analysis for each trait. As stated above, the sign on the correlation throughout all of the studies should signify the same thing. Sarah decided that positive signs indicate a positive correlation between the trait and age, and negative signs mean the opposite. See Table 2.

Step 2: Determine the Characteristics of your ES

Within a meta-analysis, all of the ES should be independent. In other words, participants cannot be repeated in different ES. If a study contains two or more ES with repeating participants (e.g., two dependent variables that are measuring similar constructs), then the ES can be averaged together to form one ES for that study. With each ES, you also need to determine how many participants it was based on.

Jacob. Jacob had three studies on gender differences and network size, with sample sizes of 30, 50, and 100.

Sarah. Sarah had two studies on Big Five traits and age. The five meta-analyses will not be independent of each other because all are based on the same group of participants. Though this lack of independence should be noted, it does not represent a violation of the basic rule of independence within a meta-analysis. The sample sizes were 80 and 200 for the two studies.

Step 3: Find your ES

An ES may be simpler to calculate for some studies compared with others. For any studies using a metric of linear correlation such as *r*, the coefficient itself is an ES. For studies comparing two means, Cohen's *d*, Hedges' *g*, or *r* can be calculated. In an ANOVA, not every ES accompanying the *F* value can be used. Specifically, an ES of an omnibus *F* value (i.e., any *F* with more than 1 *df* in the numerator) cannot be included in the meta-analysis because an omnibus effect does not test a directional (focused) hypothesis.

Regardless of whether the studies used the same analysis strategy or used different ones (e.g., one used a *t*-test, one used a correlation, and one was a chi-squared test based on a 2 × 2 frequency table), it is necessary to convert them all to the same ES metric. Formulas for conversion are readily accessible, and they are easy to calculate (Lipsey & Wilson, 2001; Rosenthal, 1991); some formulas are provided below.

Jacob. Jacob used independent *t*-tests. For each study, he can calculate Cohen's *d* using Formula 1 if he has an equal number of male and female participants.

$$d = \frac{2t}{\sqrt{df}} \tag{1}$$

If he did not have equal number of male and female participants, he can use Formula 2:

$$d = \frac{t\,(n_1 + n_2)}{\sqrt{df}\,\sqrt{n_1 n_2}} \tag{2}$$

If he performed *t*-tests but wants his ES metric to be *r*, he should use Formula 3:

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{3}$$

Jacob could also calculate *d* by the definitional formula wherein the difference between means is divided by the pooled within-group standard deviation (see Rosnow & Rosenthal, 2009). Or he could just correlate gender (as a dummy variable) with social network size and the resulting *r* is the ES.

The square of this value (r^2), as well as similar squared expressions of ES (e.g., partial eta squared or η_p^2), have some specialized uses in meta-analytic work (Rosenthal, Rosnow, & Rubin, 2000), but squared expressions of ES should not be used routinely in meta-analysis because, being squared, they no longer have the sign to indicate the directionality of the results, and they give a misleading impression of the importance and size of the effect.

Sarah. Sarah does not need to do any additional work to calculate ES because she used correlations in the first place. Each correlation coefficient is an ES.

Gordon. A third researcher, Gordon, presents a more ambiguous situation. Gordon conducted four studies on the impact of angry versus happy emotional state on skin conductance. All were repeated-measures designs in which the same participants experienced both emotional states (in a counterbalanced order). The meta-analyst has options for deriving ES in repeated-measures studies (or a matched design), which center around whether the reduction in error variance gained in the repeated-measures design (as reflected, for example, in a matched *t*-test) should be reflected in the ES or not (Dunlap, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 2002). If it is, ES will be bigger than it would be if the between-participants variance had not been subtracted out. An additional layer of complexity arises when some studies are within-participants and some are between-participants. In this case, the choices multiply but further discussion would be out of place in the present primer.

Step 4: Calculate Weighted Mean ES

In a fixed effects analysis, the weighted mean ES is calculated, with the goal of giving more weight to larger studies. The weighted mean can be misleading, however, if the studies differ

© 2016 John Wiley & Sons Ltd

in important ways methodologically and sample size is correlated with those differences. If Sarah's larger study had only men and her smaller study had only women, then the weighted mean would disproportionately represent the result for men. Therefore, if Sarah suspected or knew that sample size is confounded with a moderator variable (gender in this example), then the fixed model might be a poor choice. In the fully random effects approach we present later, studies receive equal weight.

Jacob. Although formulas are available for calculating weighted *d* (Shadish & Haddock, 2009), the simplest thing is to convert each *d* to *r* using Formula 4 (or use Formula 3 for converting *t* to *r*) and calculate the weighted mean in the same way that Sarah does (see next paragraph). It is easier to calculate an overall weighted ES using *r* than *d* if one does not have specialized software readily available.

$$r = \sqrt{\frac{d^2}{d^2 + \frac{1}{P*Q}}} \tag{4}$$

In Formula 4, *P* is the proportion of the sample in one group (e.g., the number of men divided by the total sample size) and *Q* is the proportion of the sample in the other group (e.g., the proportion of women).

Sarah. In order to calculate the weighted mean for correlations, Sarah needs to first do Fisher's *z* transformation for normalization (Fisher's *z* and the Pearson *r* will be the same for *r* values ≤ .24). This can be done through the Excel function "=fisher(x)" or through an online calculator.[4] After Fisher's *z* transformation, the ES is now represented as r_z and these can be combined meta-analytically using Formula 5 with *N* representing the number of participants contributing to a given ES.

$$Weighted\ \bar{r}_z = \frac{\sum([N-3]r_z)}{\sum(N-3)} \tag{5}$$

All calculations are done on r_z but you would convert this back to the *r* metric for presentation in a manuscript because *r* is more readily interpretable.

Step 5: Understand your ES

Once an overall mean ES is calculated for a mini meta, it can be interpreted. The conventional classification of ES is small, medium, or large, but any such designation is highly context dependent; researchers should always ask "relative to what?" rather than relying on an all-purpose standard. This could mean relative to earlier results from one's own or others' laboratories, relative to an absolute standard (e.g., variance accounted for), relative to other correlates of the same variables, relative to the general size of effects in one's subfield, and so forth. Some researchers are discouraged by small ES, but small effects can also be impressive as suggested by Prentice and Miller (1992) and are ubiquitous in psychological, medical, and public health research (Rosenthal & Rosnow, 2008).

Jacob. For Cohen's d, the magnitude of the effect also can be categorized as small ($d = .20$), medium ($d = .50$), or large ($d = .80$) (Cohen, 1988). Keep in mind, however, that what is small and large depends on many factors as mentioned above. Converting Jacob's mean ES from correlation to d using Formula 6 below yielded $d = .49$.

$$d = \frac{2r}{\sqrt{1 - r^2}} \tag{6}$$

Sarah. For correlation coefficients, magnitude can be classified as small ($r = .10$), medium ($r = .30$), or large ($r = .50$) (Cohen, 1988). In Sarah's mini metas, Extraversion and Openness to Experience had small magnitudes, and the other three traits demonstrated medium to large magnitudes by this standard.

Gordon. Interpreting the magnitude of ES presents a challenge when some of the ES in the dataset have more control of error variance than others (which is the case in repeated-measures and matched-pair designs). However, it is not one that is unique to repeated-measures and matched-pair designs, as it would apply anytime the studies vary in their error-control practices (e.g., using more homogeneous samples, using more reliable measures, or adding covariates to the analysis). In a large enough meta-analysis, one could introduce moderator variables to take such variations into account. However, this would often not be possible in a mini meta.

Step 6: Stouffer's Z Test

If the researcher wants a summary p-value for all of the studies, the Stouffer formula is simple and commonly used (Mosteller & Bush, 1954); other methods for reaching aggregate conclusions about statistical significance of the whole group of studies are available (Hedges, Cooper, & Bushman, 1992; Rosenthal, 1978). For each study's p-value, find the Z (standard normal deviate) that corresponds to it and attach the appropriate sign (which should match the sign on its respective ES), then apply Formula 7. The resulting value is itself a Z, which can be converted to its p-value. The symbol k refers to the number of independent Zs being combined.

$$Z_{combined} = \frac{\sum Z}{\sqrt{k}} \tag{7}$$

Researchers who also wish to report fixed effects confidence intervals will find guidance in Borenstein et al. (2009).

Jacob. Jacob needs to look up the Z corresponding to each p-value in his studies, using standard Z tables in any introductory statistics textbook or an online calculator and paying attention to whether each p-value was one- or two-tailed.

Sarah. Sarah, too, must find the Zs based on her p-values. She can easily calculate each Z with Formula 8 as well, based on the r and the total number of participants associated with that particular r. Again, the signs on the Z and r must match.

$$Z = r\sqrt{N} \tag{8}$$

Step 7: Random Effects Approach

The random effects approach discussed here is what we shall call the fully random effects approach. It uses statistical methods identical to those used in ordinary data analysis, with each ES treated as a data point (r_z if using the correlation metric) and the N being the number of studies involved (for illustration of this method, see DiMatteo, 2004). In the fully random effects approach, the overall effect is simply the arithmetic average of all the ES. The test of whether the average ES is greater than zero is the one-sample *t*-test, where the *N* is the number of independent ES.

In the fully random effects approach, equal weights are assigned to all ES (i.e., no weighting according to sample size). It should be noted, however, that for meta-analyses with large discrepancies in sample sizes between ES and no suspected correlations between sample size and various study characteristics, it may be preferable to use a random effects approach that does not assign equal weight to all (see Borenstein et al., 2010).

Jacob. Jacob would take the average of the three ES to find the overall magnitude of his studies. He can then enter the three ES as data points and analyze them using a one-sample *t*-test.

Sarah. Sarah, too, would average her two ES for each personality trait. She can then enter the two ES as data points and analyze using a one-sample *t*-test for each mini meta.

Optional Step 8: Heterogeneity Test

Sometimes the researcher doing a fixed effects analysis wants to ask if the ES are more variable than expected by chance. For this, a chi-squared test is done using Formula 9. With *df* equal to the number of studies minus 1, a significant result indicates more heterogeneity than expected from normal sampling variation. This test is also referred to as Q_{within} (Borenstein et al., 2009).

$$\chi^2_{k-1} = \sum (N-3)(r_z - \textit{weighted } \bar{r}_z)^2 \tag{9}$$

In Formula 9, weighted $\bar{r}_z$ is the weighted mean ES calculated in Step 4.

This method of calculating heterogeneity only informs the researcher whether heterogeneity exists (if it is significant) or not. One complementary test of heterogeneity is the I^2 index, which captures proportion of variability due to heterogeneity between studies (for formulas, see Higgins, Thompson, Deeks, & Altman, 2003). Unfortunately, both of these tests have low power when performed with few studies (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006), so please be cautious when interpreting heterogeneity in mini metas.

Optional Step 9: Contrast Analysis

Contrasts enable a person to compare ES with each other, according to moderator variables (for example, ES for male samples compared with ES for female samples, age comparisons or trends across studies, and comparing one method against another). Performing fixed effects contrasts requires a new ingredient not used in the preceding formulas: contrast weights for each study reflecting the desired comparison or trend. These must be established for each collection of studies, using standard methods for establishing contrast weights, with the constraint that the contrast weights must add to zero across studies (not just across groups of studies as defined by the moderator variable being tested). Formula 10 shows the formula, in which contrast weights are expressed as lambda (λ):

$$\textit{Contrast} = Z = \frac{\sum(\lambda * r_z)}{\sqrt{\sum\left(\frac{\lambda^2}{N-3}\right)}} \tag{10}$$

The formula yields a *Z* (standard normal deviate) that, if corresponding to a significant *p*-value, indicates that the contrast was significant. Interpretation is done by visually comparing the weighted mean ES for the studies at each level of the moderator variable.

If there are only two studies being compared, the simpler *Z* test for comparing independent correlations can be used (Formula 11):

$$Contrast\ between\ 2\ correlations = Z = \frac{r_{z\,1} - r_{z\,2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \tag{11}$$

In Formula 11, $r_{z\,1}$ refers to the Fisher's *z* transformed correlation of the first study, and N_1 is the total number of participants in the first study. Similarly, $r_{z\,2}$ refers to the Fisher's *z* transformed correlation of the second study, and N_2 is the total number of participants in the second study.

Whereas the combined *p*-value may show that one's collection of studies is collectively significant in supporting an effect, the contrast enables further assertions. A significant contrast demonstrates the presence of a moderator, while a nonsignificant contrast may have theoretical value as well (e.g., one could conclude the effect is similar in both men and women, if gender was the tested moderator). A random effects contrast can also be done, for example with a simple *t*-test between two subgroups of studies, where the ES within each subgroup have equal weighting. With few studies, statistical power for this random effects analysis is low and one's inferential strength is weak because the moderator may be confounded with other study features; this is a problem with any moderator analysis in meta-analysis, but it is especially a problem when there are only a few studies in each comparison group.

Step 10: Report your Meta-Analysis

There are no exact formats for reporting a meta-analysis, particularly a mini meta. The mini metas that we have encountered differ greatly in terms of their reporting. Some only included the mini metas within footnotes (Williams & DeSteno, 2008), while others have dedicated a whole section before the General Discussion (Case et al., 2015; Horgan, Schmid Mast, Hall, & Carter, 2004). Some authors chose to report only the mini metas without reporting results of individual studies (Hall, Roter, Milburn, & Daltroy, 1996). Others had tables showing results for each study, followed by the meta-analyzed results across studies (Hall et al., 2016; Goh et al., 2016). Regardless of your reporting choice, it is important to report what type of method you used (e.g., *d* as the ES metric, fixed effects, Stouffer test, etc.).

Jacob's summary might look like this:

> We meta-analyzed our three studies using fixed effects in which the mean effect size (i.e., mean correlation) was weighted by sample size. We first converted our Cohen's *d* into Pearson's correlation for ease of analyses. All correlations were then Fisher's *z* transformed for analyses and converted back to Pearson correlations for presentation. Overall, the effect was highly significant, $M\ r = .24$, $Z = 3.25$, $p = .001$, two-tailed, such that women have more friends than men do. A fully random effects test of the overall effect was also significant, as indicated by a one-sample *t*-test of the mean ES against zero, $M\ r = .26$, $t(2) = 7.88$, $p = .016$, two-tailed.

Sarah's summary could read:

> A separate meta-analysis was performed for each Big Five trait. We used fixed effects in which the mean effect size (i.e., mean correlation) was weighted by sample size. All correlations were Fisher's *z* transformed for analyses and converted back to Pearson correlations for presentation. Across the two studies, age was significantly positively associated with most of the Big Five traits. Neuroticism

(M $r = .45$, $p < .001$) showed the strongest effect, followed by Conscientiousness (M $r = .33$, $p < .001$), Agreeableness (M $r = .31$, $p < .001$), and Extraversion (M $r = .22$, $p < .001$), all two-tailed. Openness to Experience (M $r = -.13$, $p < .10$) did not reach significance using the Stouffer's Z test. None of the five one-sample t-tests against zero (random effects approach) yielded a significant result (not in table), which is not surprising considering only two studies contributed to each of the tests.

Limitations

There are several limitations pertaining to the current primer as well as conducting mini metas in general. In our attempts to make mini metas as accessible as possible, we could mention only a few meta-analytical procedures. Obviously, we could not cover all of the contingencies of research design or statistical analyses, and there are procedures we did not cover because they are not relevant to an internal meta-analysis (estimating publication bias, for example) or not possible with only a few studies (such as multiple regression). We recommend interested readers to seek alternative resources for other meta-analytic procedures or more detailed descriptions of the current procedures outlined in this primer (for books, see Borenstein et al., 2009; Cooper et al., 2009; Cumming, 2013; Lipsey & Wilson, 2001).

While we argued that mini metas can encourage a more open science, they are nevertheless still susceptible to questionable research practice (Ueno, Fastrich, & Murayama, 2016). In particular, when doing a mini meta, researchers can still employ biased stopping rules such that they check the significance level after each replication (as opposed to running the number of predetermined replications then conduct a mini meta afterwards) or exclude studies in order to yield significant results in their internal meta-analyses. As Ueno and colleagues argued, this flexible stopping rule can increase Type I error rates. This is certainly true if a researcher is adamantly focusing on the p-values obtained from a mini meta. However, we strongly encourage researchers to focus more on the meta-analytic ES rather than the corresponding p-value. Furthermore, as Schmidt (1996) and many others have pointed out, research designs and practices protect an investigator from Type I error far better than they protect from Type II error. Reducing Type II error is one of the large advantages of doing meta-analysis.

Conclusion

This primer covered some of the basics for calculating a meta-analysis for a handful of studies within a manuscript. We hope it will be useful and will inspire more researchers to conduct meta-analyses no matter how many or how few studies they have on hand.

Short Biographies

Jin X. Goh is a doctoral student in social psychology and personality at Northeastern University. His research focuses on the relationship between prejudice and nonverbal behavior. He is also interested in accuracy and bias in social perception. He holds a BA in psychology from Bard College.

Dr Judith A. Hall is University Distinguished Professor of Psychology at Northeastern University in Boston, MA. She has dual, and overlapping, interests in nonverbal communication processes and in quality of medical care with a focus on physician–patient communication. Themes in her research have included gender and social power in relation to nonverbal communication, and the measurement and correlates of accuracy in person perception. She has been editor in chief of *Patient Education and Counseling* and the *Journal of Nonverbal Behavior*, and is currently an associate editor at that journal. Dr Hall has been an author or editor of several books on nonverbal behavior, interpersonal accuracy, and physician–patient communication. She received her PhD in Social Psychology from Harvard University and has held positions at the Johns Hopkins University and the Harvard School of Public Health.

Dr Robert Rosenthal's substantive research interests are in the areas of (a) interpersonal expectancy effects, including the effects of experimenters' expectations on the results of their research and the effects of teachers' expectations on their students' intellectual development, and (b) the operation of different channels of nonverbal communication. His methodological research interests include (a) experimenter-sources and participant-sources of artifact in behavioral research, (b) significance testing in scientific research, and (c) a number of quantitative procedures including contrast analysis and meta-analysis. He received the SPSP 2015 Methodological Innovation Award for his numerous contributions to psychology. He holds a BA (1953) and a PhD (1956) both in Psychology from UCLA. He taught for 5 years at the University of North Dakota, 37 years at Harvard University, and has been at the University of California, Riverside, for 17 years.

Notes

* Correspondence: Department of Psychology, Northeastern University, 360 Huntington Ave., Boston, MA 02115. Email: jin.x.goh@gmail.com

[1] Readers should be reminded of the bidirectional nature of confidence intervals around effect sizes whether they are significant or not. For any given nonsignificant effect, the weight of evidence that the true effect is zero is no greater than the weight of evidence that the effect is dramatically larger than the obtained effect (Rosenthal & Rubin, 1994).

[2] The annotated Excel spreadsheet can be found on the Supplementary page, our lab website www.northeastern.edu/socialinteractionlab/publications/, OSF page osf.io/6tfh5/, or by contacting the first author. Check these websites for updates.

[3] For alternative meta-analytic procedures that are not restricted by fixed or random effects, see varying coefficient models (Bonett, 2008).

[4] There are several online calculators available for Fisher's z transformation (and transforming back to Pearson's correlation). An example is http://onlinestatbook.com/analysis_lab/r_to_z.html

References

Bonett, D. G. (2008). Meta-analytic interval estimation for Pearson correlations. *Psychological Methods*, **13**, 173–189.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). *Comprehensive Meta-analysis (version 3).* Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-analysis.* United Kingdom: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods,* **1**, 97–111.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, **9**, 333–342.

Carter, J. D., Hall, J. A., Carney, D. R., & Rosip, J. C. (2006). Individual differences in the acceptance of stereotyping. *Journal of Research in Personality*, **40**, 1103–1118.

Case, C. R., Conlon, K. E., & Maner, J. K. (2015). Affiliation-seeking among the powerless: Lacking power increases social affiliative motivation. *European Journal of Social Psychology*, **45**, 378–385.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin,* **112**, 155–159.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The Handbook of Research Synthesis and Meta-analysis* (2nd edn), (pp. 357–376). New York, NY: Russell Sage Foundation.

Cumming, G. (2013). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, **25**, 7–29.

DiMatteo, M. R. (2004). Social support and patient adherence to medical treatment: A meta-analysis. *Health Psychology*, **23**, 207–218.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, **1**, 170–177.

Goh, J. X., Schlegel, K., Tignor, S. M., & Hall, J. A. (2016). Who is interested in personality? The Interest in Personality Scale and its correlates. *Personality and Individual Differences*, **101**, 185–191.

Hall, J. A., Blanch, D. C., Horgan, T. G., Murphy, N. A., Rosip, J. C., & Mast, M. S. (2009). Motivation and interpersonal sensitivity: Does it matter how hard you try? *Motivation and Emotion*, **33**, 291–302.

Hall, J. A., Goh, J. X., Schmid Mast, M., & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, **84**, 433–445.

Hall, J. A., Roter, D. L., Milburn, M. A., & Daltroy, L. H. (1996). Patient's health as a predictor of physician and patient behavior in medical visits: A synthesis of four studies. *Medical Care*, **34**, 1205–1218.

Hall, J. A., Ship, A. N., Ruben, M. A., Curtin, E. M., Roter, D. L., Clever, S. L., Smith, C. C., & Pounds, K. (2015). Clinically relevant correlates of accurate perception of patients' thoughts and feelings. *Health Communication,* **30,** 423–429.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin*, **111**, 188–194.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, **327**, 557–560.

Horgan, T. G., Schmid Mast, M., Hall, J. A., & Carter, J. D. (2004). Gender differences in memory for the appearance of others. *Personality and Social Psychology Bulletin*, **30**, 185–196.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index?. *Psychological Methods*, **11**, 193–206.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, **15**, 342–345.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, **23**, 524–532.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … & Nosek, B. A. (in press). *Reducing Implicit Racial Preferences: II.* Intervention effectiveness across time. *Journal of Experimental Psychology:* General.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, **4**, 863.

Lamarche, V. M., & Murray, S. L. (2014). Selectively myopic? Self-esteem and attentional bias in response to potential relationship threats. *Social Psychological & Personality Science*, **5**, 786–795.

Lim, D., & DeSteno, D. (2016). Suffering and compassion: The links among adverse life experiences, empathy, compassion, and prosocial behavior. *Emotion*, **16**, 175–182.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, **26**, 1827–1832.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical Meta-analysis.* Thousand Oaks, CA: Sage publications.

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, **9**, 343–351.

McNemar, Q. (1962). *Psychological Statistics*, 3rd edn. New York: Wiley.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, **7**, 105–125.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology: Volume I. Theory and Method*. Cambridge, MA: Addison-Wesley.

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin*, **142**, 79–106.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, **112**, 160–164.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, **85**, 185–193.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, **86**, 638–641.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research (rev. ed.).* Newbury Park, CA: Sage Publications.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to Nonverbal Communication: The PONS Test.* Baltimore: The Johns Hopkins University Press.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of Behavioral Research* (3rd edn). New York: McGraw-Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach.* New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, **5**, 329–334.

Rosnow, R. L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Journal of Psychology*, **217**, 6–14.

Rule, N. O., Tskhay, K. O., Brambilla, M., Riva, P., Andrzejewski, S. A., & Krendl, A. C. (2015). The relationship between anti-gay prejudice and the categorization of sexual orientation. *Personality and Individual Differences*, **77**, 74–80.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115–129.

Shadish, W. R., & Haddock, C. K. (2009) Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analysis* (2nd edn) (pp. 257–277). New York, NY: Russell Sage Foundation.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, **143**, 534–547.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, **145**, 643–654.

Vazire, S. (2016). Editorial. *Social Psychological & Personality Science*, **7**, 3–7.

Williams, L. A., & DeSteno, D. (2008). Pride and perseverance: The motivational role of pride. *Journal of Personality and Social Psychology*, **94**, 1007–1017.

Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin*, **142**, 165–197.