Remedies for three common issues in meta-analysis

Daniel S. Quintana

NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and
Addiction, University of Oslo, and Oslo University Hospital, Oslo, Norway

Corresponding author: Daniel S. Quintana, email: daniel.quintana@medisin.uio.no

Meta-analysis has become a popular tool to synthesise data from a body of work investigating a common research question. Combining several effect sizes facilitates the estimation of a summary effect size. As this provides a putatively objective means of summarising a body of data, it has become the most cited type of study design (Patsopoulos, Analatos, & Ioannidis, 2005) and the gold standard for evidence when developing health guidelines (Harbour & Miller, 2001). However, there are several issues in meta-analysis that can contribute to inaccuracies. Due to these limitations, meta-analysis — or “mega-silliness”, to borrow a term from an early detractor (Eysenck, 1978) — has been criticised for almost as long as the method has been available. There are several tools to help ensure meta-analysis quality, which will be outlined later in this article. But first, let’s consider this consider this (fictitious) extract from the results section of a meta-analysis that I will use to briefly illustrate three common issues in meta-analysis:

“As hypothesised, a meta-analysis synthesising eight effect sizes from seven studies suggests that social media usage is positively associated with self-reported levels of anxiety [$r = .073$, 95% CI ($.004, 0.142$); $p = .038$; Fig. 1A]. Egger’s regression test suggested no evidence of publication bias ($p = .19$; Fig. 1B).”

When this imaginary author — who also happens to be releasing a book on the dangers of social media — initially searched for studies, they actually found nine studies fulfilling their inclusion criteria. However, the summary effect size of this meta-analysis of nine studies [$r = .062$, 95% CI ($-.004, 0.127$)] was associated with a p-value of .061 (Fig. 2). As this outcome would hinder the chances of publication (and book sales), they took a closer look at study 6, which reported a negative association between social media usage and anxiety levels. This study happened to be the earliest study included in the meta-analysis. As this study did not include Instagram usage (which became popular during 2012) they adjusted the eligibility criteria to only include studies published from 2013

Figure 1. The relationship between social media usage and self-reported anxiety. The forest plot (a) visualises point estimates for a set of hypothetical studies with filled squares. The diamond reflects the estimated summary effect size. The funnel plot (b) illustrates each effect size against its standard error.

onwards, justified by a (post hoc) hypothesis that Instagram usage is most associated with anxiety. With the removal of the study, the meta-analysis became statistically significant. While this fabricated scenario might seem extreme at first glance, almost 40% of over 2,000 surveyed psychologists admitted that they had excluded data after looking at the impact of doing so (John, Loewenstein, & Prelec, 2012).

Meta-analysis pre-registration provides a remedy to analytic flexibility as it provides an *a priori* record of the researcher's analysis intentions. In the above example, exclusion criteria would have been specified before the analysis was conducted. A common critique of pre-registered analyses is that they offer no flexibility for analysis. However, this is incorrect as pre-registration does not lock the researcher into a specific plan but instead helps clearly delineate pre-planned and exploratory analysis. In the above case, the researcher would be free to report their analysis with the removal of the study, but they would have to state their reasoning, which is flimsy at best. For guidance, the Preferred Reporting Items for Systematic Reviews and Meta-Analysis protocols (PRISMA-P) document provides a list of items that should be addressed in a meta-analysis protocol (Shamseer et al., 2015).

There are several options available for meta-analysis pre-registration. First, a time-stamped meta-analysis protocol following PRISMA-P guidelines can be uploaded to Open Science Framework

Figure 2. The original set of studies for the assessment of the relationship between social media usage and self-reported anxiety. Upon realising that the removal of study 6 would result in a statistically significant result, our fictitious researcher removed this study using a post hoc justification. Meta-analysis pre-registration would have prevented a post hoc adjustment of exclusion criteria dressed up as being *a priori*.

(<http://osf.io>). Second, if the meta-analysis investigates a health-related outcome, it can be pre-registered at PROSPERO (<https://www.crd.york.ac.uk/prospERO/>). The PROSPERO website guides the researcher through a step-by-step registration form, which is used to generate a pre-registration record. These records can be amending, but any changes are timestamped. Third, a meta-analysis protocol can be submitted for peer-review at a number of journals, such as Systematic Reviews. While the peer-review process obviously takes more time than uploading a protocol, it can provide important feedback that can improve the meta-analysis. Fourth, a meta-analysis can be submitted using the Registered Report format, in which the meta-analysis is reviewed in two stages. In the first stage, the meta-analysis introduction and methods sections are peer-reviewed before data collection. If the study rationale and methodological approach is deemed appropriate, the journal will provide provisional acceptance of the final paper. Once the data is collected and analysed, the paper is submitted again for the second stage review which includes with the same introduction and methods as stage one, but with the addition of results and conclusions. In the second stage of review, the focus is on whether the results and interpretation are in line with the pre-approved method. Importantly, paper acceptance is not contingent on the statistical significance of results. As of August 2018, there are 42 journals that

Figure 3. A comparison of conventional meta-analysis funnel plots vs. contour-enhanced funnel plots. Compared to the original funnel plot (a), which is also presented in Fig 1B, a contour enhanced funnel plot (b) visualises ranges of statistical significance. The gray-shaded region corresponds to p -values between .10 and .05, whereas the dark gray-shaded region corresponds to p -values between .05 and .01.

offer Registered Report meta-analysis (<https://cos.io/rr/>), with the majority of these journals publishing research from the biobehavioral sciences.

Publication bias is a well-recognised issue in meta-analysis (Sterne, Gavaghan, & Egger, 2000). As statistically significant studies are more likely to be published, it is likely that there are several non-significant studies left languishing in researcher's file-drawers for any given research area, which cannot contribute to meta-analyses. Thus, a meta-analysis summary effect size may not best represent the "true" effect size. A common approach to assess the risk of publication is to construct a funnel plot, which plots effect sizes against a measure of variance (e.g., standard errors). A symmetrical funnel plot is indicative of low risk of publication bias (Egger, Smith, Schneider, & Minder, 1997). Despite the inherent subjectivity of visually inspecting a funnel plot to assess evidence of asymmetry, many researchers still rely on this tool.

Now let's return to our imaginary study. As our fictitious researcher was aware of issues with subjective funnel plot assessments, they objectively assessed asymmetry by constructing a funnel plot. As researchers are poor at identifying asymmetry from funnel plots (Terrin, Schmid, & Lau, 2005), they followed recommendations and performed Egger's regression test (Egger et al., 1997) to complement the funnel plot (Fig. 1A). A statistically significant test is indicative of funnel plot asymmetry. In this case, the test was not statistically significant ($p = .19$). Thus, our researcher

concluded that there was no risk of publication bias. A closer look at the data, however, reveals that this conclusion was most likely incorrect.

A common misunderstanding of funnel plots and their associated tests is that they measure publication bias, when in fact they measure *small study* bias (Sterne et al., 2011). This type of bias might include publication bias, but it also encompasses other types of bias commonly associated with small studies, such as study quality. A relatively simple funnel plot modification, called a contour-enhanced funnel plot, can help better understand the risk of publication bias compared to small study bias. For a given range of effect sizes and variances, it is possible to calculate levels of statistical significance at every point. As a result, it is possible to visualise ranges of statistical significance (e.g., $p = .05$ to $.01$) on a funnel plot. Comparing our original funnel plot (Fig. 3A) with a contour-enhanced funnel plot (Fig. 3B) reveals that despite no apparent evidence of funnel plot asymmetry, half of the included effect sizes had p -values were very close to $.05$. It is highly improbable that half of the available p -values would lie so close to $.05$. Either these significant studies were also the result of selective analyses or there are several unpublished studies that exist. Several measures have developed to correct for bias in meta-analysis, such as p -curve (Simonsohn et al., 2014), p -uniform (van Aert, Wicherts, & van Assen, 2016), and the three-parameter selection model (Iyengar & Greenhouse, 1988; McShane et al., 2016), with the latter demonstrating superior performance (Carter et al., 2018). But without access unpublished research these methods will always be estimates, at best. One plausible solution to this scenario is to conduct a meta-analysis that only includes pre-registered empirical studies, such as the recent meta-analysis on “power posing” from Gronau and colleagues (2017).

Now to our final common issue in meta-analysis: effect size dependence. Notice that in the earlier meta-analysis example there were eight effect sizes but only seven studies? This was due to two effect sizes being extracted from study 4 (Fig. 1A). This is a common occurrence as papers sometimes report several outcome measures. However, as these effect sizes were derived from the same study population they are statistically dependent. As typical meta-analysis procedures assume that effect

sizes are statistically independent, combining dependent meta-analyses can lead to biased results (Hunter & Schmidt, 2004).

There are several approaches to for addressing statistical dependency in meta-analysis. First, you can simply choose one effect size per study. While straightforward, a pre-specified system is required to help select the chosen effect sizes, otherwise it can be tempting to simply choose the largest effect. What's more, this approach excludes potentially valuable effect sizes from your analysis. To include more than one effect from a study, it is possible to statistically aggregate effects (Hunter & Schmidt, 2004). But to do this, you need to include a measure of within-study correlation, which is almost never reported in papers. If the paper contains raw data, an occurrence which is almost as rare, within-study correlations can be calculated. Therefore, researchers almost always need to estimate within-study correlations if this approach is to be taken. Alternatively, robust variance estimation can be used to account for effect size dependency without a need to know within-study correlations (Hedges, Tipton, & Johnson, 2010).

Meta-analysis is an important tool for better understanding often disparate research areas. This method can provide a more objective summary of a research area than simply “vote counting” the number of significant and non-significant results. But like almost any statistical tool, meta-analysis can facilitate inaccurate inferences if used inappropriately. With this article, I hope to have provided a brief overview highlighting some common issues in meta-analysis and practical solutions to these problems. For an example meta-analysis that has been pre-registered, appropriately treats small-study bias, and accounts for dependent effect sizes using summary statistics, see Murray and colleagues (2018). The R script and data to perform the analyses described in the present paper are available at <https://osf.io/3vjm5/>.

References

- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2018, July 31). Correcting for bias in psychology: A comparison of meta-analytic methods. <https://doi.org/10.31234/osf.io/9h3nu>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical Research Ed.)*, 315, 629–634.
- Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, 33(5), 517.
- Gronau, Q. F., Erp, S. V., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology*, 2(1), 123–138. <https://doi.org/10.1080/23743603.2017.1326760>
- Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *BMJ: British Medical Journal*, 323(7308), 334–336.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Murray, S. B., Quintana, D. S., Loeb, K. L., Griffiths, S., & Le Grange, D. (2018). Treatment outcomes for anorexia nervosa: A systematic review and meta-analysis of randomized controlled trials. *Psychological Medicine*, 1-10. doi:10.1017/S0033291718002088
- Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. A. (2005). Relative citation impact of various study designs in the health sciences. *JAMA*, 293(19), 2362–2366. <https://doi.org/10.1001/jama.293.19.2362>

Shamseer, L., Moher, D., Clarke, M., Gherzi, D., Liberati, A., Petticrew, M., ... Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ (Clinical Research Ed.)*, 349, g7647.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6), 666–681.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53, 1119–1129.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., ... Schmid, C. H. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj*, 343, d4002.

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, 58(9), 894–901. <https://doi.org/10.1016/j.jclinepi.2005.01.006>

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting Meta-Analyses Based on p Values: Reservations and Recommendations for Applying p-Uniform and p-Curve. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 11(5), 713–729. <https://doi.org/10.1177/1745691616650874>