Meta-Analysis of Social Science Research: A Practitioner's Guide

Zuzana Irsova[1], Hristos Doucouliagos[2], Tomas Havranek[3], and T. D. Stanley[4]

Journal of Economic Surveys, forthcoming, 2024

Abstract

This article provides concise, nontechnical, step-by-step guidelines on how to conduct a modern meta-analysis, especially in social sciences. We treat publication bias, *p*-hacking, and systematic heterogeneity as phenomena meta-analysts must always confront. To this end, we provide concrete methodological recommendations. Meta-analysis methods have advanced notably over the last few years. Yet many meta-analyses still rely on outdated approaches, some ignoring publication bias and systematic heterogeneity. While limitations persist, recently developed techniques allow robust inference even in the face of formidable problems in the underlying empirical literature. The purpose of this paper is to summarize the state of the art in a way accessible to aspiring meta-analysts in any field. We also discuss how meta-analysts can use advances in artificial intelligence to work more efficiently.

Keywords: meta-analysis, publication bias, *p*-hacking, artificial intelligence, model uncertainty

JEL Codes: A14, B49, C83

[1] Zuzana Irsova, Institute of Economic Studies, Faculty of Social Sciences, Charles University, Prague, and Anglo-American University, Prague. Irsova acknowledges support from the Czech Science Foundation (#23-05227M). Data and codes (R and Stata) for modern meta-analyses can be found at meta-analysis.cz.

[2] Hristos Doucouliagos, Department of Economics, Deakin University, Melbourne, VIC 3125, Australia.

[3] Tomas Havranek, Institute of Economic Studies, Faculty of Social Sciences, Charles University, Prague; Centre for Economic Policy Research, London; and Meta-Research Innovation Center at Stanford. Havranek acknowledges support from the Czech Science Foundation (#21-09231S) and from NPO "Systemic Risk Institute" LX22NPO5101, funded by the European Union - Next Generation EU (Czech Ministry of Education, Youth and Sports, NPO: EXCELES).

[4] Corresponding author: T.D. Stanley, Department of Economics, Deakin University, Melbourne, VIC 3125, Australia. stanley@hendrix.edu

I. INTRODUCTION

Meta-analysis has grown into a thriving research industry. According to Google Scholar, more than 107,000 meta-analyses were published in 2022 alone.[1] Even in economics, which had long been skeptical, meta-research is now published (e.g., Andrews & Kasy 2019, Brodeur et al. 2020, Brown et al. 2023, DellaVigna & Linos 2022, Elliot et al. 2022, Havranek et al. 2024, Neisser 2021) and cited (e.g., Angeletos & Huo 2021, Cogley & Jovanovic 2022, Comin et al. 2021, Kroodsma et al. 2018, List & Uhlig 2017) in the most august journals, and meta-analyses often represent the most cited studies for individual prestigious outlets.

Over the last several years, we have seen important advancements in methods and the rigor in which typical meta-analyses are conducted. Nonetheless, in our roles as editors and reviewers, we also see many meta-studies fall short in applying appropriate statistical analysis. Because our field has been quite dynamic, it is understandable that some researchers have fallen behind. We all struggle to keep pace in our respective fields of expertise. Although the current state of art in meta-analysis methods and its consequences have been shared at conferences, seminars, and referee reports, we believe that it is a propitious time to briefly summarize in practical and nontechnical terms what is widely accepted as best practice.

These guidelines, of course, cannot be the final word on how meta-analysis should be conducted. Meta-analysis is a complex and rapidly evolving field, and context together with newly developed approaches may force researchers to deviate from these or any set of guidelines. Yet we believe that many researchers, especially those with modest meta-analysis experience, will benefit from following the guidelines. They reflect a distillation of methodological contributions across economics, psychology, and medical research, and also our experience in applying these meta-analysis methods widely across disciplines and hundreds of specific areas of research. Although our focus is on economics and related disciplines, we believe that these guidelines are sufficiently general to be helpful to *any* meta-analysis. These methods guidelines are intended to complement the existing reporting guidelines for meta-analysis published in this *journal* (Havranek et al., 2020). Together, they form the natural starting point for any aspiring meta-analyst – though she will also do well to consult the other existing "how to" protocols (e.g., Borenstein et al. 2021, Gurevitch et al. 2018, Higgins et al. 2022, Koricheva et al. 2017, Moher et al 2015, Nakagawa et al. 2017, Nakagawa et al. 2023, Page et al. 2021).

Meta-analysis, if failing to use up-to-date methods, can be as misleading as a good meta-analysis is enlightening to policymakers and researchers. An especially important issue is publication selection bias and *p*-hacking. Out of the 107,000 meta-analyses published in 2022, slightly more than half do not discuss publication bias at all.[2] Because publication bias or *p*-hacking can easily exaggerate the typical reported effect size by a factor of 2 or more (Ioannidis et al. 2017, Open Science Foundation 2015, Camerer et al. 2018; Bartoš et al. 2023b), meta-analyses that ignore publication bias may potentially cause more harm than good. Many advanced techniques for publication selection bias correction with rigorous foundations have recently been introduced and supported by Monte Carlo simulations and dozens of applications. Other recent developments include the treatment of observed and unobserved systematic heterogeneity in the context of model uncertainty and some forms of *p*-hacking. Together, these method advances constitute important steps forward in the understanding and interpreting of contemporary research.

We start by discussing the search for primary studies to be included in meta-analysis. Then we move to data collection, the treatment of publication bias and heterogeneity, and, lastly, the estimation of conditional meta-analysis means corrected for both publication bias and systematic methodological problems found in some primary studies (misspecifications). Before concluding the paper, we provide a short, bullet-point checklist. The website meta-analysis.cz contains many examples of modern meta-analyses together with their data and codes for R and Stata. For ease of exposition we speak directly, in the second person, to aspiring meta-analysts.

II. LITERATURE SEARCH

You should conduct meta-analyses only on topics you understand thoroughly. That is, you should have conducted primary research on the topic, written a detailed narrative literature review, or taught extensively on the subject. If not, you will need a co-author from this specific sub-field. If a meta-analysis on the topic already exists, you must show a strong *raison d'être:* why does your meta-analysis add value? The lack of accommodation for publication bias or heterogeneity in the original meta-analysis is such a reason. The fact that several new primary studies have been published does not. You need to show, at a minimum, a substantial advance in the methods that you use in comparison to the original meta-analysis. Mechanical updates should be left as training exercises for undergraduate students or artificial intelligence. Of course, exceptions to this general advice should be made when there have been important advancements in the approaches and/or methods of this particular area of research, placing the robustness of past meta-analysis results into question. In addition, structural features of societies might have changed such that previous effect sizes might not be representative. In some fields, it is sensible to test the influence of cutoff years to see whether structural breaks are present (for example, the Great Recession or the Covid-19 pandemic).

Now, based on your knowledge of the topic, assemble a list of 5 primary studies that you certainly must include in the meta-analysis. To ensure that you have selected the 5 most important studies, you may enlist a large language model. But be careful about relying too much on artificial intelligence since current models often provide results that are factually incorrect; always double-check and give priority to your expertise. (Useful guidelines for employing artificial intelligence in the context of economics education and research are provided by Cowen & Tabarrok 2023.) Then design your main search query in Google Scholar. We prefer Google Scholar to other databases because it includes all papers that have appeared online and goes through the full text of papers, not just the title, abstract, and keywords. Having one main query for one universal database helps other researchers replicate your meta-analysis. Note, however, that Google Scholar's algorithms may change at any moment. Depending on your topic, there can be value in using an additional database (such as EconLit for economics), and it might increase the robustness of your approach. Use different combinations of the keywords employed in primary studies. You will know that your query is reasonably well prepared if the 5 most important primary studies identified above are among the first hits. Spend several days fine-tuning the query (that is, improving the percentage of highly relevant primary studies returned among the first 50 hits) and pay attention to the correct search syntax. For inspiration, see the "examined studies" section in the online appendix to Havranek (2015): meta-analysis.cz/eis.

For modern meta-regression analysis (MRA) techniques to work, you need at least 30 estimates of the effect size reported in at least 10 primary studies.[3] Ideally, you will end up with many more. Your Google Scholar search will return hundreds of studies. Read the abstracts of the first 500 of them and download all that could potentially contain empirical estimates of the effect you are interested in. Go through the downloaded studies in detail recording all reported estimates of the effect in question and their standard errors (or measures from which the standard errors can be computed, such as *p*-values and *t*-statistics). Standard errors (SEs) are typically needed for weights and publication bias correction. However, in some specific literatures, standard errors may not be commonly reported, and sample sizes can serve as substitute for standard errors, as SEs are often approximately proportional to 1/√n. In fact, it can be argued that using inverse sample sizes (or degrees of freedom) in place of SEs is superior when correlations or partial correlations are the effect sizes meta-analyzed (Hunter & Schmidt 1990, Stanley & Doucouliagos 2012; 2023). In other cases of missing SEs, individual primary studies may report dozens of effect size estimates (sensitivity analyses or scenarios). Some meta-analyses (Havranek et al. 2015a, Matousek et al. 2022) have used this within-study dispersion to approximate study-level confidence and, from this, bootstrapped study-level standard errors. Within-study dispersion should be treated as a last resort, explicitly acknowledged in the paper and used along with robustness checks that employ subsets of the literature with reported standard errors.

Do not exclude any study *ex ante* because you suspect the study is of poor quality, or because it is published in a local journal. You can always conduct subsample analysis in which you show what happens when you exclude some studies. In general, you want to include all studies that meet minimum explicitly-stated inclusion criteria, because they allow you to identify how variations in methodology affects the results – indeed, that might be your main reason for conducting the meta-analysis. The weight you place on bad studies may (and often will) be close to zero, but the decision should be carefully justified in your meta-analysis. Similarly, do not omit unpublished studies. While the inclusion of unpublished studies by itself is unlikely to solve publication bias, there can be systematic differences between published and unpublished studies. What if you have too many eligible primary studies, perhaps hundreds, more than you can feasibly collect? The best option here is to invite co-authors who help you collect the entire dataset, excluding no primary study. If adding co-authors is impossible, you may need to use a random subset of the literature or to limit your analysis to a scientifically meaningful and well-defined subset. Using a random subset is also a last resort that you should avoid if possible and fully reveal when employed.

Next, do ‘snowballing.’ You already have primary studies you are sure you will use. Gather their references (for example, using Scopus or Web of Science) and inspect the 100 studies that are most commonly cited among the primary studies identified in your Google Scholar search. This way you can be reasonably confident you have not missed any important primary study. Of course, you can never be sure you have included all available studies. In particular, new studies will have few citations, so will not typically appear among the first hits in Google Scholar, nor will they be identified via the snowballing approach described above. You should repeat your Google Scholar search but limit it just for the last three years. Then inspect the abstracts of the first 30 hits. You should also inspect recent citations (those from the last three years) for the three most important primary studies. Be sure to make notes during the entire literature search process to facilitate replicability and construct a PRISMA diagram (read Havranek et al. 2020, Moher et al. 2015, Page et al. 2021, for details). See meta-analysis.cz/frisch (Elminejad et al. 2022a) or meta-analysis.cz/risk (Elminejad et al. 2022b) for an example of the diagram.

III. DATA COLLECTION

You and your co-authors should collect data for meta-analysis; the task cannot be delegated to research assistants. Perhaps in a few years artificial intelligence (GPT 7?) will be able to help with this laborious task, but, for now, we see no substitute to the authors of the meta-analysis, experts on the meta-analyzed literature, carefully going through the primary studies one by one and painstakingly creating their dataset by hand, one data point after another. In fact, as noted by the philosopher and economic historian Deirdre McCloskey (McCloskey 2016), here we should not talk about *data* (“things given” in Latin), but *capta*: “things seized.” Unlike the authors of most econometric studies, meta-analysts do not take existing data but create new databases. Examples of meta-analysis datasets are available at meta-analysis.cz.

If possible, at least two co-authors should collect the data independently. The reason is that random mistakes in manual coding of studies (dozens of pages in pdf) are inevitable, and with two experts collecting the same data the mistakes can be easily identified and corrected. The effect sizes collected for meta-analysis must be comparable quantitatively, not only qualitatively. This means that not only the same estimated sign implies an effect in the same direction, but that it is meaningful to compare the actual size of effects across primary studies. Quantitatively comparable effect sizes include correlation coefficients, odd ratios, elasticities, dollar values, and standardized mean differences. Regression coefficients are generally not comparable quantitatively without further transformations, because different primary studies can use different units of measurement or functional forms of the independent and/or dependent variables. An exception is represented by regressions in which variables on both sides are used in logarithms and, therefore, the regression yields estimated of elasticities. If the authors of primary studies report summary statistics for their regression variables, the results can often be recomputed to a common metric. To take one example, the effect of class size on student achievement can be gauged by the change in the average test score, measured in percentiles of the test score's standard deviation, in response to an increase in class size by ten students (Opatrny et al. 2023).

If such standardization is infeasible, meta-analysts can recompute regression estimates to partial correlation coefficients (Doucouliagos 2005, Zigraiova & Havranek 2016, Cazachevici et al. 2020). However, a lot of information is lost though this transformation as well as the practical interpretation of the original effect sizes. Partial correlations should thus be used as a *last* resort (Stanley & Doucouliagos 2012, Stanley & Doucouliagos 2023, Roth et al. 2018). If you use partial correlations in your main analysis, always include a robustness check that focuses on the largest subset of primary studies with comparable effect sizes (often elasticities). For similar reasons, we discourage the use of simple correlation coefficients in meta-analysis if more informative alternatives, such as standardized mean differences, are available. Doucouliagos (2011) provides preliminary guidelines for interpreting partial correlations by trying to map partial correlations to elasticities. Because partial correlations (and, for that matter, all correlations) are related to their standard errors by construction (Stanley et al. 2023b), it is often a good idea to transform them to Fisher's z statistics for analysis, and then transform them back to correlations. Alternatively, you may use the instrumental variable approach by Irsova et al. (2023).

You should collect all estimates reported in the primary studies. There are three good reasons for doing so. First, if you want to put more weight on the estimates preferred by authors themselves, you can always present a meta-analysis of the corresponding subsample of the dataset. This restricted analysis can serve as a robustness check or the baseline but does not justify discarding information by completely ignoring other estimates. Second, primary studies often report robustness checks themselves, and sometimes these estimates are deemed inferior by the authors. By using all estimates, you can examine whether the “inferior” estimates systematically differ from those preferred by authors. In your final best-practice meta-analysis estimate, you can still put more weight on the authors’ preferred results. Third, sometimes it is simply impossible to judge *objectively* which estimates the authors prefer. Collecting and analyzing all estimates eliminates the need for such a judgement.

Primary studies typically report the standard errors of the estimated effect sizes. If instead *t*-statistics or *p*-values are reported, standard errors can be easily computed from these quantities. Complications arise in regression analysis if the explanatory variable of interest is included as an interaction with another variable or is included in different functional form (for example quadratic). Then, sometimes, it is straightforward to compute the corresponding effect size as the partial derivative of the estimated regression with respect to the explanatory variable of interest evaluated at the sample mean. But the issue is more challenging for the computation of the standard error, and the delta method needs to be used (Oehlert 1992, Liu 2012). Because data on covariances are almost never reported in primary studies, meta-analysts typically use the delta method with the assumption of zero covariances. An example of a dataset where the delta method is used is available at meta-analysis.cz/spillovers (Havranek & Irsova 2011).

Note that meta-analysis can be conducted also for graphical results, not just numerical ones. In that case meta-analysts need to carefully convert graphs to numbers using pixel coordinates (Ehrenbergerova et al. 2023, Fabo et al. 2021, Havranek & Rusnak 2013, Rusnak et al. 2013); a concrete example of graphical data collection is available at meta-analysis.cz/house_prices/IRs.pdf. Measurement error is inevitable when coding graphical results, but it is comparable to rounding in the case of numerical results – perhaps even less problematic, because the measurement error for coding graphical results is likely to be random.

You should carefully inspect outliers and influence points in your data. Construct a funnel plot (a scatter plot of effect sizes and their precision). If some data points are far away from the main funnel shape (Egger et al. 1997, Stanley 2005) or raise a red flag in DFBETA (a useful method for measuring the influence of individual data points on regression analysis, Belsley et al. 1980), read again the corresponding primary studies to make sure there are no typos in your data or in the primary study itself. Perhaps, further careful reading will identify some nuance in the way the study was conducted that makes its results not quantitatively comparable to the rest of the research literature. If still in doubt, write to the authors of the primary study. Perhaps reported units or your understanding of them are wrong. Influence or leverage points, as identified by DFBETA, are especially important as they can have a lot of weight and bias your meta-analysis results. Thus, these need to be corrected or, as a last resort, removed. Report robustness checks on what happens when you drop the outliers or when you winsorize (that is, replace observations above and below a certain centile with the value of that centile; Bajzik et al. 2021, Zigraiova et al. 2021) the data. The point is that your results should not be driven by a small number of highly influential research findings unless you know them to be especially reliable studies, in which case you must justify their prominence in detail.

Apart from effect sizes and standard errors, you should also collect information on the main differences in the context in which the estimated effect sizes were obtained. Most meta-analyses should collect at least 10 variables (often dummy binary variables which take the value 0 or 1) that reflect differences in data, methods, and publication characteristics, commonly many more depending on the size and complexity of the database, but we encourage meta-analysts to keep the number below 30 for parsimony. For example, does the experiment in the primary study focus on a representative sample of the population, or only on the elderly? In which country was it conducted? Was a placebo or an alternative treatment assigned to the control group? When was the study published, what is the impact factor of the outlet, and how many per-year citations has the study received?

Some researchers have argued that measures of publication impact reflect a ‘winner’s curse’, where the most highly cited papers and journals tend to be the most highly exaggerated (Ioannidis 2005; Young et al. 2008; Costa-Font et al. 2013). However, some reviewers may demand that the meta-analyst evaluate research quality by these conventional metrics. In some contexts, the number of citations or the impact factor of the outlet may serve as proxies for unobservable quality characteristics. While variables related to publication can be used in a similar form in almost every meta-analysis, the remaining variables will vary. Meta-analysts should carefully prepare a list of variables they need to code before they start actual data collection. This is perhaps the most difficult and creative part of a meta-analysis: the number of potential variables is almost unlimited, and you must select the most important ones based on previous discussions in the literature and your own expertise. A large language model can be useful to help identify some of the dimensions in which the primary studies vary. Again, be aware that artificial intelligence can provide misleading results. Always double-check.

You may want to include additional information that complements what you collect from primary studies. For example, if the primary studies were conducted using data from many different countries, it can be a good idea to include country (or region) characteristics as additional variables in meta-analysis. The results of an experiment can be influenced by temperature or humidity, and the response of inflation to interest rate hikes can depend on the financial development of the country. In this way meta-analysis can bring further value added and insight, often impossible to analyze by the individual primary studies.

IV. PUBLICATION BIAS AND *P*-HACKING

A key issue that is almost impossible for individual primary studies to address is publication bias and *p*-hacking. That is, in contrast to what has sometimes been suggested (e.g., Rothstein et al. 2005, Rothstein, 2008), publication bias is not a problem *of* meta-analysis. It is a problem of primary empirical research, and meta-analysis represents one of two ways of effectively addressing the bias. Preregistration of large multi-lab experiments is the other (Nosek et al. 2018, Klein et al. 2014, 2018). When preregistration is fully followed, insignificant results are unlikely to be hidden in a file drawer nor will authors be tempted to *p*-hack their data or methods in order to provide significant results. Preregistration is less likely to work well in observational research, where researchers can inspect their data before preregistration. In contrast, meta-analysis can be used to correct for publication bias under all circumstances, with or without preregistration, when enough primary studies have been conducted on the specific research question. Still, current bias-correction techniques are not perfect, and a combination of pre-registration and meta-analysis seems ideal to combat publication bias even in the case of observational research.

Definitions of publication bias and *p*-hacking vary. Sometimes the former is defined generally to comprise all situations in which the observed research results do not correspond to the results authors obtain when they analyze their data for the first time. Sometimes publication bias refers only to a situation where some studies are unpublished (the ‘file drawer’ problem) because their results are insignificant or unintuitive. With the narrower definition of publication bias, *p*-hacking denotes conscious or unconscious manipulation of data or methods until statistical significance is achieved. In practice, both phenomena are observationally equivalent to the meta-analyst (unless nontraditional data are available, see Brodeur et al. 2023), so the broader definition of publication selection bias often encompasses both. But publication bias and *p*-hacking, narrowly defined, may have different implications for correction methods.

If *p*-hacking is extreme enough, no publication bias correction can succeed. Consider, for example, the hypothetical case in which many researchers are dishonest and unscrupulous, make up their data, and cheat with estimation results. Then anything is possible in the research record, and meta-analysis will fail. But nothing suggests we live in such a world, and fraud, when influential on the meta-analysis findings, can sometimes be discovered (e.g., via DFBETA) and omitted – though of course by far not all influential observations are necessarily fraudulent. Comparisons of preregistered replications and original research results suggest an exaggeration of reported results due to publication bias and *p*-hacking (Kvarven et al. 2020), but there is little evidence of widespread or outright cheating. Journals have increasingly required data and codes for published papers, which should reduce or eliminate the more extreme forms of *p*-hacking (Askarov et al. 2023). As long as *p*-hacking is limited to selecting samples, outcome measures, and estimation methods technique to achieve statistical significance in a preferred direction, meta-analysis can accommodate and greatly reduce the resulting bias.

As of 2023, we find it indefensible to ignore publication bias and *p*-hacking in a meta-analysis, unless meta-analysis is used to summarize the findings from multi-lab replications or individual patient data. As we noted in the Introduction, more than half of all meta-analyses published in 2022 unfortunately do ignore publication bias, often simply reporting fixed-effect or random-effects estimates and stopping there. In our view, such summaries without further correction convey little information. Of course, there are important exceptions. For randomized controlled trials of a few brand-new medical treatments or other interventions, a simple summary of current best evidence may be sufficient to guide policy and to indicate where further advancement may be made. In many areas of experimental research where there are only a handful of studies, it may be impractical to go beyond simple weighted averages.

IV.1 MAIN APPROACHES

If you want to report a simple summary statistic before moving to a more sophisticated analysis, you should opt for unrestricted weighted least squares (UWLS), which dominate both fixed-effect and random-effects meta-analysis estimators (Stanley & Doucouliagos 2015, 2017; Stanley et al. 2023a). Likewise, it is never enough to use one arbitrary test of publication bias and say that because the test does not reject the null hypothesis of no bias, you will ignore bias and *p*-hacking in the rest of the analysis. You should use several approaches, or a Bayesian model average across them (Bartoš et al. 2023a), and always discuss the bias-corrected estimates even if you somehow reject the presence of bias.

There are two broad method families of bias correction techniques. One family is based on selection models (van Assen et al 2015, van Aert & van Assen 2021, Andrews & Kasy 2019, Hedges 1984, 1992, Iyengar & Greenhouse 1988, Vevea & Hedges 1995) which assume that estimates with different significance levels have different probabilities of publication. These models are typically estimated by maximum likelihood and can be interpreted as re-weighting the observed estimates by the inverse publication probability. The second family of techniques is based on the funnel plot (Bom & Rachinger 2019, Duval & Tweedie 2000, Egger et al. 1997, Furukawa 2019, Ioannidis et al. 2017, Stanley 2008, Stanley & Doucouliagos 2012, Stanley & Doucouliagos 2014) and assumes that selective reporting works via the size of the reported estimate (instead of the *p*-value, as selection models assume). Funnel-based techniques estimate the effect size reported in a hypothetical, infinitely precise study. Both groups of models have their pros and cons, and you should use, at least as a robustness check, models from both families. We prefer funnel-based techniques, because they are more flexible and can also incorporate some forms of *p*-hacking, not just publication bias, as we will soon see.

Among selection models, the one with the most rigorous foundations is Andrews & Kasy (2019). You should also report the results of a simplified selection model, *p*-uniform* (van Assen et al 2015, van Aert & van Assen 2021), which can be more stable under some circumstances (McShane et al. 2016, van Aert & Niemeyer 2022, Irsova et al. 2023). Among funnel-based techniques, the baseline is PET-PEESE (Stanley & Doucouliagos 2014), which has been found to work best among bias-correction techniques when compared to preregistered replications (Kvarven et al. 2020). Another model, endogenous kink (Bom & Rachinger 2019), improves the performance of PET-PEESE in some situations. A useful robustness check is provided by WAAP (Ioannidis et al. 2017), which focuses on the estimates that are adequately powered. Codes for these techniques are available at meta-analysis.cz under the heading "new papers". The meta-analysis of Havranek et al. (2024) on the elasticity of substitution between skilled and unskilled labor, presents an example of up-to-date application of these techniques, and can serve as a practical template.

A recent alternative to the above application of multiple methods is to use a Bayesian model average, RoBMA-PSMA, across them (Bartos et al. 2023a, Maier et al. 2023). RoBMA-PSMA is a sophisticated weighted average over both families of models that uses the full research record to calculate the weights. There is also a tutorial for RoBMA-PSMA that employs a menu-driven program, JASP, complete with its own instructional video (https://bit.ly/pubbias), that does these complex calculations for you (Bartos et al. 2022). Also, JASP has drop-down menu choices that calculate: selection models, PET-PEESE, WAAP, *p*-curve, and *p*-uniform.

IV.2 IMPORTANT DETAILS

Note that if you collect more than one estimate per study (which we recommend, because using just one estimate means that you ignore a lot of information), you need to make two adjustments. First, include a robustness check that additionally weights each estimate by the inverse of the number of estimates reported per study. The adjustment is easy to implement in meta-regression estimators such as PET-PEESE and endogenous kink. No easy adjustment exists for selection models, and the *p*-uniform* model can only be conducted using one estimate per study – typically the median estimate. It may be important in practice whether equal weight is placed on each estimate or each study, depending on which of the two can be viewed as the natural unit of analysis. If individual studies mostly use one study-specific dataset to deliver various estimates obtained using different methods, the study is the natural unit of analysis; if individual studies often examine different subsamples of data or entirely different datasets within studies, estimates are the natural units of analysis. For example, Krueger (2003) shows that, in the literature on the effect of class size on student achievement, the two approaches give substantially different results.

Second, you should cluster standard errors at the study level. The clustering option is again easy to implement in meta-regression models, and the Andrews & Kasy (2019) model also allows for clustering. But clustering works poorly when the number of clusters is small or when clusters are uneven —that is, if some studies report many estimates but others only a few. If you have fewer than 40 studies, you should use wild bootstrap instead (Roodman et al. 2019, used in the applications of Gechert et al. 2022 at meta-analysis.cz/sigma or Yang et al. 2023 at meta-analysis.cz/hedge) or the approach of Pustejovsky and Tipton (2022). Note that the clustering of standard errors at the study level does not fully address potential estimate dependence caused, for example, by sample overlap. Sample overlap is an important issue, and we recommend the solution suggested by Bom and Rachinger (2020). An imperfect remedy is two-way clustering at the level of studies and countries or datasets (Havranek and Irsova, 2017). In addition, with more than one estimate per study meta-regression methods can (and, at least as a robustness check, should) include study-level dummies to filter out unobserved study-level heterogeneity that might be correlated with the publication bias term. This adjustment can automatically be accomplished by fixed-effects panel models, for example, in STATA. Note, however, that using study-level dummies means that you eliminate between-study variation and rely on within-study variation. This reliance will be problematic if studies report only a small number of estimates or if observed standard errors contain measurement error caused by, for example, rounding as fixed-effects panel models are more susceptible to attenuation bias than standard meta-regression.[4]

All techniques mentioned above try to address publication bias. But only funnel-based techniques can additionally address some forms of *p*-hacking; selection models assume that reported results are individually unbiased (Mathur & VanderWeele 2020), which is incompatible with any *p*-hacking. If the authors of primary studies *p*-hack their effect size estimates in response to the precision given by their data and methods in order to obtain statistically significant results, funnel-based methods like PET-PEESE and endogenous kink still come close to recovering the underlying ‘true’ effect size. Because all abovementioned estimators rely on inverse variance weighting, they may have problems if the reported precision is also substantively *p*-hacked.

In addition, funnel-based techniques detect publication bias through a correlation between the reported effect size and its standard error that is caused by truncation when there is selection for statistical significance (Stanley & Doucouliagos 2017). However, medical researchers argue that this correlation could arise due to some unspecified ‘small-study’ effects. We routinely deal with this potential conflicting interpretation by controlling for any systematic heterogeneity through MRA. Another answer to questions about ‘small-study’ effects is to use a new test, PSST (proportion of statistical significance test), that does not depend, in any way, on a correlation of SE (or sample size) with effect size (Stanley et al. 2021). PSST has been shown to be more powerful than selection models and funnel-based methods in detecting publication bias should it exist.

Irsova et al. (2023) present a new estimator, MAIVE, that is based on PET-PEESE and the seminal idea of Stanley (2005) published in this *journal*. MAIVE takes the square root of the inverse of the sample size of primary studies as an instrument for reported precision and can thus address publication selection on estimates and/or their standard errors. MAIVE is also useful in other situations in which estimates are correlated with standard errors (e.g., when the meta-analysis includes correlations, Cohen’s *d*, or an inversion of the original regression estimate: see Stanley & Rosenberger 2009, Havranek et al. 2024).

The instrumental approach is especially suitable if you suspect that some method choices in primary studies can jointly affect both the estimated effect size and the standard error. By using an instrument for the standard error, ideally also with study-level dummy variables (or fixed-effects panel) in the regression, you control for unobserved heterogeneity that might otherwise contaminate your analysis of publication bias and *p*-hacking. MAIVE is therefore a useful robustness check, though it remains to be seen whether *p*-hacking on standard errors is important in practice; likely it is much less common than *p*-hacking on effect size estimates. In any case, the problem can be addressed in funnel-based models by using the instrumental variable approach, while no such a straightforward solution exists for selection models. An R package for MAIVE is available at meta-analysis.cz/maive.

V. HETEROGENEITY AND IMPLIED ESTIMATES

Few empirical literatures can be represented by a single mean estimate, even when corrected for publication bias and taking unobserved heterogeneity into account. You should examine observed systematic heterogeneity; that is, examine why individual reported estimates of effect sizes vary.

Eventually, the goal is to provide implied estimates, conditional means of effect sizes for different scenarios reflecting different contexts in which the effect size can be estimated or for which policy may be especially relevant. In the discussion of data collection, we have already mentioned that, if possible and permitted by the size of the database, you need to code at least 10 variables that capture the most important features of data, method, and publication characteristics of estimates and studies. For many meta-analyses, you will need to code many more. It will not hurt to ask an artificial intelligence (AI) program (e.g., chatGPT) to help identify the most important dimensions in which primary research studies on the topic differ, as long as you use your own professional judgment as the final arbiter. As we have noted earlier, you should always double-check the outputs of large language models. See meta-analysis.cz/sigma (Gechert et al. 2022) or meta-analysis.cz/eis (Havranek 2015) for specific examples how a final dataset with many variables capturing heterogeneity looks and which variables are sensible to code and collect.

There are two ways to approach observed heterogeneity in meta-analysis. The first way is to repeat the procedure described in the previous section about publication bias for various subsets of the dataset, the subsets driven by the main variables believed to capture heterogeneity. For example, studies can be divided according to countries, methods, or data age. As a result, you will get conditional estimates for various empirical contexts. The advantage of the subset approach is that quite different studies, and indeed quite different effect sizes, can be summarized in one paper through separate subgroup meta-analyses. At some point, when two groups of studies are different enough to warrant a separate subset analysis, they should remain separate. Of course, exceptions are possible, and subset analysis is always useful as a robustness check for multiple meta-regression if you are unsure.

The second way to address heterogeneity is multiple MRA where the heterogeneity variables are included (together with the standard error) on the right-hand side of the regression model and the estimated effect sizes define the dependent variable. You should treat MRA as an extension of PET-PEESE. If you have good empirical reasons to doubt the performance of PET-PEESE regarding publication bias (including *p*-hacking) in your specific case (for example, an instability of central estimated coefficients resulting for small changes in the regression method or model), you will want to put more weight on the subset analysis mentioned above.

The multiple meta-regression approach has two key advantages over subset analysis: it is relatively parsimonious, allowing inference from a single specification (unlike of many distinct subsets), and it accounts explicitly for likely omitted-variable bias in observational primary studies as well as in the MRA itself. Variables that reflect heterogeneity are often correlated and investigating them in isolation can easily lead to biased results. Nevertheless, this advantage is also related to an important limitation of multiple MRA. With many explanatory variables that are correlated among themselves, collinearity arises, and the resulting individual meta-regression coefficient estimates are imprecise and hard to interpret. In addition, with multiple meta-regression you face model uncertainty: you do not know *ex ante* which variables to include in the final model. If you include all that you have collected, chances are that many will prove irrelevant and/or redundant which will again increase the imprecision of the entire MRA results.

A solution that tackles both model uncertainty and collinearity is Bayesian model averaging with a dilution prior (George 2010, Eicher et al. 2011, Steel 2020). Bayesian model averaging runs many regressions with different combinations of right-hand-side variables and weights them according to data fit and model complexity. The dilution prior adds a weight that penalizes models with high collinearity. This model ensemble has been successfully employed in many meta-analyses (Bajzik et al. 2021, Elminejad et al. 2022a, 2022b, Havranek et al. 2024, among others), and an example of the code is available at meta-analysis.cz/students/students.do.

If you want to avoid Bayesian approaches, you can use frequentist model averaging (Hansen 2010, Amini & Parmeter 2012), which has less frequently been applied in meta-analysis (Kroupova et al. 2022, for example). Frequentist model averaging addresses model uncertainty but not collinearity, so you must carefully inspect variance-inflation factors and remove (or merge) variables with the factor above 10.

In general, as in primary data analysis, you will not be able to use binary variables (dummy variables which take the values of either 0 or 1) that show little variance – often those with means below 0.03 or above 0.97 because, when included in a regression model, such variables typically create problems with collinearity and can lead to volatile results for the entire model. These variables should be avoided even when you use Bayesian model averaging. If your data has little collinearity, you may also use less complex techniques, such as the general-to-specific approach

(Efroymson 1960, Smith 2018), in which the least significant variables are gradually eliminated prior to estimating the final model. In general, it is a good idea to use at least two of the three aforementioned approaches (Bayesian averaging, frequentist averaging, general-to-specific), one as the baseline and another as a robustness check.

Which weights should you use for multiple meta-regression? Here again we recommend robustness checks. The optimal meta-analysis weight is based on inverse variance, but in multiple MRA it can potentially lead to a level of collinearity that defeats the original purpose of making the estimation more efficient. A discussion of the pros and cons of various weights is available in Zigraiova & Havranek (2016). You should use the classical inverse-variance weight as the starting point. If you have a strong reason to be concerned about collinearity, your model averaging specification can also be unweighted (Matousek et al. 2022) or weighted by the inverse of the number of estimates reported per study, which gives each study the same weight (Havranek et al. 2018c). Collinearity is an issue only if you need a reliable estimate of the effect of specific variables that are highly correlated with others. For overall 'prediction' and best practice, collinearity typically does not matter. Again, we recommend estimating at least two of these models, one as the baseline, another as a robustness check. Should you have concerns about *p*-hacking on the standard error, you may marry the instrumental (MAIVE) and Bayesian model averaging approaches (Strachan & Inder 2004, Koop et al. 2012), though, to our knowledge, such an approach has not so far been used in meta-analysis – so this is low-hanging fruit for technically skilled meta-analysts.

As the central culmination of your meta-analysis, you should provide conditional means of estimated effect sizes for different scenarios. For subset analysis, the derivation of conditional means is straightforward, as we have already noted. For multiple MRA you need to compute fitted (or predicted) values from the estimated meta-regression. That is, you plug in specific values for right-hand-side variables and recover the implied effect size on the left-hand side. To make this exercise feasible, you will need to define a baseline 'best practice' in the literature, or several versions of best practice when there is ambiguity. For example, we prefer studies that use the strongest available methodology: randomized experiments and quasi-experimental designs when available, controls for endogeneity when relevant, panel models rather than cross-sectional or time series data, and studies that omit the fewest relevant control variables. Note that the resulting estimate is corrected for publication bias (and many forms of *p*-hacking), approximately, by substituting zero for the standard error variable. The definition of 'best practice' is, to some degree, unavoidably subjective, but it can be, aside from the meta-analyst's expertise, based on a recent and highly regarded primary study. For examples and more discussion of conditional means and best practice, see Bajzik et al. (2020, meta-analysis.cz/armington), Havranek et al. (2024, meta-analysis.cz/skill), or Cala et al. (2023, meta-analysis.cz/incentives).

VI. CHECKLIST: HOW TO DO A MODERN META-ANALYSIS

1) Choose a topic you or your co-authors know well from your own primary research.

2) Choose a topic for which no prior meta-analysis exists. If you update a meta-analysis, you need to use new and stronger methods.

3) Prepare a search query in Google Scholar. Inspect the first 500 hits.

4) Inspect the 30 studies that are most cited among the ones included based on the Scholar search.

5) Do not discard any study *a priori* based on publication outlet or perceived quality.

6) Collect all estimates and their standard errors, when possible, not just one estimate per study.

7) Collect the data independently with a co-author, then compare and correct mistakes.

8) Use original effect size measures when comparable. If not, transform them to a common metric.

9) Correlations (including partial ones) should be used as a last resort.

10) Inspect outliers and influence points but be careful about deleting or winsorizing them. Report robustness checks.

11) Think carefully about the aspects in which primary studies differ. Collect at least 10 variables capturing this heterogeneity.

12) If you want to report a simple summary statistic, use the unrestricted weighted least squares weighted average, rather than fixed-effect or random-effects estimates.

13) Always correct for publication bias (including *p*-hacking). Use RoBMA-PSMA or at least one technique from each of the following model groups: selection models (Andrews and Kasy, *p*-uniform*), funnel-based models (PET-PEESE, endogenous kink, WAAP, MAIVE).

14) Report standard errors clustered at the study level. With fewer than 40 studies use the wild bootstrap.

15) If possible, in meta-regressions use study-level dummy variables (i.e., fixed-effects panel models) to filter out unobserved study-level heterogeneity.

16) Estimate the multiple meta-regression model by applying Bayesian model averaging with the dilution prior.

17) If collinearity is not at issue, also use frequentist model averaging or the general-to-specific approach.

18) Provide conditional means for effect sizes in different situations (corrected for both publication bias and potential method weaknesses in some studies).

Of course, there are important exceptions that will depend on practical considerations and the complexities of the specific area of research investigated to this or to any sparse imperative checklist. We see these guidelines as a useful starting point, not as the final word about conducting meta-analyses.

VII. CONCLUSION

Meta-analysis methodology has improved dramatically over the last few years, leading the charge towards a credibility revolution in the social sciences and beyond. Recent advances include solutions to: *p*-hacking, model uncertainty, collinearity, and to the lack of robustness in earlier approaches to publication bias correction. Yet few applied meta-analyses have fully exploited these advances. The purpose of this paper is to summarize these recent advances, along with providing straightforward practical guidelines for conducting meta-analysis, and to do so in one brief, nontechnical document accessible to meta-analysts from different fields.

As of 2023, meta-analysis provides much more than a weighted average of the existing empirical literature. For one, neither primary studies nor weighted averages, alone, can account for publication bias and *p*-hacking. Moreover, as we have discussed, meta-analysis can bring substantial value added by including external information: for example, linking regional and institutional characteristics to the results of primary studies conducted for different countries (Havranek et al. 2015b, Havranek & Irsova 2017, Havranek et al. 2018b). Meta-research can also identify and measure the impact of potential method problems in some studies, such as endogeneity (Valickova et al. 2015, Havranek et al. 2018a, Kroupova et al. 2022) or attenuation bias (Havranek et al. 2024) – problems that are, again, difficult to tackle in individual primary studies without a systematic comparison to the rest of the literature.

By necessity, this brief sketch is incomplete in its breath, depth, and nuance. It is offered as a starting point for those new to meta-analysis and as a concise discussion of central methodological issues facing the meta-analysis of economics and the social sciences. Sensible deviations in our specific recommendations are welcome, especially from researchers with experience and/or strong statistics/econometric backgrounds. Nonetheless, we feel strongly that all meta-analyses should use methods that explicitly deal with common issues found in social science research: publication selection bias (including *p*-hacking), systematic heterogeneity, and the dependence of multiple estimates within studies. Although we believe that many disciplines and areas of research could benefit from these suggestions, we recognize that not all are suitable for multiple regression. With a median of 5 estimates/study (Stanley et al. 2023), most of systematic reviews of medical research are not sufficiently informative for MRA. However, much of social science research could benefit from the routine use of meta-regression, rather than subgroup comparisons, as long as there are approximately 10 or more estimates per coded moderator variable.

Going forward, we see two important issues for meta-research. The first one is increased openness and transparency. You should always provide your data and codes online. Consider uploading early, private versions of your data on the Open Science Framework, where it can be time stamped, and sharing it publicly. A potential benefit of providing your materials online early is that they make other researchers more likely to cite your work, especially if your method is novel in some ways and your code is well documented, easy to run and follow.

Secondly, the field is likely to be radically changed by artificial intelligence soon. As in any research area, the most important steps in meta-analysis are creative; thus, it is hard to imagine how these can be fully automated even with radically better versions of AI than we have at present. But meta-analysis is based on a uniquely laborious data collection that often takes months of expert researcher time. So, meta-analysis can benefit from AI more than most research fields. We believe that in a few years new versions of GPT (or some equivalent) will be able to assist with data collection from primary studies. Within a few years, AI may truly become a "virtual co-author," scraping text as a starting point, and helping to identify relevant papers, variables, and data errors.

AI programs such as GPT will soon be able to update existing meta-analyses that provide their data because this is a relatively mechanical task. GPT can be trained on the data of the original meta-analysis, the original search query, and the texts of the original primary studies and then update the dataset by scrapping data from the texts of new primary studies using best-practice meta-analysis methods in combination with these or other guidelines as a template. In most cases, therefore, it will be enough to publish one *good* meta-analysis on each empirical research topic. Updates could happen automatically, perhaps in real time as cumulative meta-analyses (Lau et al. 1992, Wetterslevet al. 2008, Kulinskaya & Mah 2022). When all authors of the original meta-analysis will provide code (or a chatGPT query) in an online appendix, readers can obtain updates with a few 'clicks.' Only major breakthroughs in methodology will warrant a new meta-analysis.

Automation due to advances in AI will enable meta-analysts to devote more of their time to the most creative parts of research, which will again increase the average quality and contribution that meta-analysis makes to collective scientific knowledge. Having undergone a period of steady and notable advancement, meta-analysis and meta-research can now lead researchers towards a broader credibility revolution in the social and medical sciences.

References

van Aert R.C. & M. van Assen (2021): “Correcting for publication bias in a meta-analysis with the p-uniform* method.” Tilburg University & Utrecht University working paper. doi: 10.31222/osf.io/zqjr9.

van Aert, R.C.M. & H. Niemeyer (2022): “Publication bias.” In: O'Donohue, W., Masuda, A., Lilienfeld, S. (eds) Avoiding Questionable Research Practices in Applied Psychology. Springer, Cham. doi: 10.1007/978-3-031-04968-2_10.

Amini, S.M. & C.F. Parmeter (2012): “Comparison of model averaging techniques: Assessing growth determinants.” Journal of Applied Econometrics 27(5): pp. 870–876.

Andrews I. & M. Kasy (2019): “Identification of and correction for publication bias.” American Economic Review 109(8): pp. 2766–2794.

Angeletos G.-M. & Z. Huo (2021): “Myopia and anchoring.” American Economic Review 111(4): pp. 1166–1200.

Askarov, Z., Doucouliagos A, and Doucouliagos H. & Stanley, T.D. (2023). “The significance of data-sharing policy.” Journal of the European Economic Association 21(3): pp. 1191–1226.

van Assen, M.A., van Aert, R. & J.M. Wicherts (2015): “Metaanalysis using effect size distributions of only statistically significant studies.” Psychological Methods 20: pp. 293–309.

Bajzik, J., Havranek, T., Irsova, Z. & J. Schwarz (2020): “Estimating the Armington elasticity: The Importance of Study Design and Publication Bias.” Journal of International Economics 127: art. 103383.

Bartoš F., Maier M., Quintana, D.S. & Wagenmakers E.J., et al. (2022). “Adjusting for publication bias in JASP and R: Selection models, PET-PEESE, and robust Bayesian meta-analysis.” Advances in Methods and Practices in Psychological Science 5(3): pp. 1–19.

Bartoš, F., Maier, M., Wagenmakers, E.J., Doucouliagos, H. & T.D. Stanley (2023a): “Robust Bayesian meta-analysis: Model averaging across complementary publication bias adjustment methods.” Research Synthesis Methods 14(1): pp. 99–116.

Bartoš F., Maier M., Wagenmakers E.J., et al. (2023b). “Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics.” arXiv:2208.12334

Belsley, D.A., Kuh, E. & R.E. Welsch (1980): "Regression diagnostics: Identifying influential data and sources of collinearity." John Wiley & Sons.

Bom, P.R.D. & H. Rachinger (2019): "A kinked meta-regression model for publication bias correction." Research Synthesis Methods 10(4): pp. 497–514.

Bom, P.R.D. & H. Rachinger (2019): "A generalized-weights solution to sample overlap in meta-analysis." Research Synthesis Methods 11(6): pp. 812–832.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & H. Rothstein (2021): "Introduction to meta-analysis." Second edition, Chichester: Wiley.

Brown A.L., Imai T., Vieider F.M. & C.F. Camerer (2023): "Meta-Analysis of empirical estimates of loss aversion." Journal of Economic Literature, forthcoming.

Brodeur, A., Cook, N. & A. Heyes (2020): "Methods matter: P-hacking and causal inference in economics." American Economic Review 110(11): pp. 3634–3660.

Brodeur, A., Carrell, S., Figlio, D. & L. Lusher (2023): "Unpacking p-hacking and publication bias." American Economic Review, forthcoming.

Cala, P., Havranek, T., Irsova, Z., Matousek, J. & J. Novak (2022): "Financial Incentives and Performance: A Meta-Analysis of Economics Evidence." CEPR Discussion Papers 17680, The Centre for Economic Policy Research, London. Available online at meta-analysis.cz/incentives.

Camerer, C., Dreber, A., Holzmeister, F. et al. (2018): "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." Nature Human Behaviour 2: 637–644.

Cazachevici, A., Havranek, T. & R. Horvath (2020): "Remittances and economic growth: A meta-analysis." World Development 134: art. 105021.

Comin, D., Lashkari, L. & M. Mestieri (2021): "Structural change with long-run income and price effects." Econometrica 89(1): pp. 311–374.

Costa-Font, J, McGuire, A. & Stanley, T.D. (2013) "Publication selection in health policy research: The winner's curse hypothesis." Health Policy, 109: 78–87.

Cowen, T. & A.T. Tabarrok (2023): "How to Learn and Teach Economics with Large Language Models, Including GPT." Available online at ssrn.com/abstract=4391863 (March 2023).

Cogley, T. & B. Jovanovic (2022): “Structural breaks in an endogenous growth model.” Review of Economic Studies 89(2): pp. 666–694.

DellaVigna, S. & E. Linos (2022): “RCTs to scale: Comprehensive evidence from two nudge units.” Econometrica 90(1): pp. 81–116.

Doucouliagos, H. (2011): “How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics.” Working Papers 5/2011, Deakin University.

Duval S. & R. Tweedie (2000): “Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis.” Biometrics 56(2): pp. 455–463.

Efroymson, M.A. (1960): “Multiple regression analysis.” In: Ralston A, Wilf HS, editors. Mathematical methods for digital computers. New York: Wiley.

Egger, M., Smith, G. D., Schneider, M. & C. Minder (1997): “Bias in meta-analysis detected by a simple, graphical test.” British Medical Journal 315(7109): pp. 629–634.

Ehrenbergerova, D., Bajzik, J. & T. Havranek (2023): “When does monetary policy sway house prices? A meta-analysis.” IMF Economic Review 71: pp. 538–573.

Eicher, T.S., Papageorgiou, C. & A.E. Raftery (2011): “Default priors and predictive performance in Bayesian model averaging, with application to growth determinants.” Journal of Applied Econometrics 26: pp. 30–55.

Elliott, G., Kudrin, N. & K. Wuthrich (2022): “Detecting p-Hacking.” Econometrica 90(2): pp. 887–906.

Elminejad, A., Havranek, T., Horvath, R. & Z. Irsova (2022a): “Publication and Identification Biases in Measuring the Intertemporal Substitution of Labor Supply.” Charles University, Prague, available online at meta-analysis.cz/frisch.

Elminejad, A., Havranek T. & Z. Irsova (2022b): “Relative risk aversion: A meta-analysis.” Charles University, Prague, available online at meta-analysis.cz/risk.

Fabo, B., Jancokova, M., Kempf, E. & L. Pastor (2021): “Fifty shades of QE: Comparing findings of central bankers and academics.” Journal of Monetary Economics 120: pp. 1–20.

Furukawa C. (2019): “Publication Bias under Aggregation Frictions: Theory, Evidence, and a New Correction Method.” Massachusetts Institute of Technology working paper. www.jeameetings.org/2019s/Gabst/1161.pdf

Gechert, S., Havranek, T., Irsova, Z. & D. Kolcunova (2022): "Measuring capital-labor substitution: The importance of method choices and publication bias." Review of Economic Dynamics 45: pp. 55–82.

George, E.I. (2010): "Dilution priors: Compensating for model space redundancy." In IMS Collections Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown, volume 6, pp. 158-165. Institute of Mathematical Statistics.

Gurevitch, J., Koricheva, J., Nakagawa, S. & G. Stewart (2018): "Meta-analysis and the science of research synthesis." Nature 555: pp. 175–182.

Hansen, B.E. (2007): "Least squares model averaging." Econometrica 75(4): pp. 1175–1189.

Havranek, T. (2015): "Measuring intertemporal substitution: The importance of method choices and selective reporting." Journal of the European Economic Association 13(6): pp. 1180–1204.

Havranek, T., Herman, D. & Z. Irsova (2018b): "Does daylight saving save electricity? A Meta-Analysis." Energy Journal 39(2): pp. 35–61.

Havranek, T. & Z. Irsova (2011): "Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect Is." Journal of International Economics 85(2): 234–244.

Havranek, T. & Z. Irsova (2017): "Do Borders Really Slash Trade? A Meta-Analysis." IMF Economic Review 65(2): pp. 365–396.

Havranek, T., Irsova, Z., Laslopova, L. & O. Zeynalova (2024): "Publication and Attenuation Biases in Measuring Skill Substitution." The Review of Economics and Statistics, forthcoming. doi: 10.1162/rest a 01227.

Havranek, T., Irsova, Z., Janda, K. & D. Zilberman (2015a): "Selective Reporting and the Social Cost of Carbon." Energy Economics 51: pp. 364–406.

Havranek, T., Irsova, Z. & O. Zeynalova (2018c): "Tuition Fees and University Enrolment: A Meta-Regression Analysis." Oxford Bulletin of Economics and Statistics 80(6): pp. 1145–1184.

Havranek, T., Horvath, R., Irsova, Z. & M. Rusnak (2015b): "Cross-Country Heterogeneity in Intertemporal Substitution." Journal of International Economics 96(1): pp. 100–118.

Havranek, T., Irsova, Z. & T. Vlach (2018a): "Measuring the Income Elasticity of Water Demand: The Importance of Publication and Endogeneity Biases." Land Economics 94(2): 259–283.

Havranek, T. & M. Rusnak (2013): “Transmission Lags of Monetary Policy: A Meta-Analysis.” International Journal of Central Banking 9(4): pp. 39–76.

Havranek, T., Stanley T.D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W.R., Rost, K. & R.C.M. van Aert (2020): “Reporting Guidelines for Meta-Analysis in Economics.” Journal of Economic Surveys 34(3): pp. 469–475.

Hedges L.V. (1984): “Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences.” Journal of Educational Statistics 9: pp. 61–85.

Hedges, L.V. (1992): “Modeling Publication Selection Effects in Meta-Analysis.” Statistical Science 72(2): pp. 246–255.

Higgins, J.P.T, Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. & V.A. Welch – editors (2022). Cochrane Handbook for Systematic Reviews of Interventions version 6.3. Cochrane, 2022. Available online at training.cochrane.org/handbook (updated February 2022).

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Ioannidis JPA (2005) “Contradicted and initially stronger effects in highly cited clinical research.” JAMA 294: 218–228.

Ioannidis, J., Stanley, T. & H. Doucouliagos (2017): “The Power of Bias in Economics Research.” Economic Journal 127(605): pp. 236–265.

Irsova, Z., Bom, P.R.D., Havranek, T. & H. Rachinger (2023): “Spurious Precision in Meta-Analysis.” MetaArXiv 3qp2w, Center for Open Science. Available online at meta-analysis.cz/maive.

Iyengar, S. & J.B. Greenhouse (1988): “Selection Models and the File Drawer Problem.” Statistical Science 3(1): pp. 109–117.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R., ..., & B. A. Nosek (2014): “Investigating variation in replicability: A ´many labs´ replication project.” Social Psychology 45(3): pp. 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … , Zhija, Z. & B. A. Nosek (2018): Many Labs 2: Investigating Variation in Replicability Across

Samples and Settings. Advances in Methods and Practices in Psychological Science 1(4): pp. 443–490.

Koop, G., Leon-Gonzalez, R. & R. Strachan (2012): "Bayesian model averaging in the instrumental variable regression model." Journal of Econometrics 171(2): 237–250.

Koricheva J, Gurevitch J. & K. Mengersen (2017): "Handbook of meta-analysis in ecology and evolution." Koricheva J, Gurevitch J. & K. Mengersen (eds.), Princeton: Princeton Univesity Press.

Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block ,B.A., Woods, P., Sullivan, B., Costello, C. & B. Worm (2018): "Tracking the global footprint of fisheries." Science 359(6378): pp. 904–907.

Kroupova, K., Havranek, T. & Z. Irsova (2022): "Student Employment and Education: A Meta-Analysis." CEPR Discussion Papers 16550, The Centre for Economic Policy Research, London. Available online at meta-analysis.cz/students.

Krueger, A. (2003): "Economic Considerations and Class Size," Economic Journal 113(485): pp. 34–63.

Kvarven, A., Stromland, E. & M. Johannesson (2020): "Comparing meta-analyses and preregistered multiple-laboratory replication projects." Nature Human Behavior 4: pp. 423–434.

Kulinskaya, E. & E.Y. Mah (2022): "Cumulative meta-analysis: What works." Research Synthesis Methods 13(1): pp. 48–67.

Lau, J., Antman, E., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & T. Chalmers (1992): "Cumulative meta-analysis of therapeutic trials for myocardial infarction." The New England Journal of Medicine 327(4): pp. 248–254.

List, J. & H. Uhlig (2017): "Introduction to The Past, Present, and Future of Economics: A Celebration of the 125-Year Anniversary of the JPE and of Chicago Economics." Journal of Political Economy 125(6): pp. 1723–1727.

Liu, X. (2012): "Survival Analysis: Models and Applications." First edition. Higher Education Press, John Wiley & Sons, Ltd.

Maier, M., Bartos, F. & E.J. Wagenmakers (2023): “Robust Bayesian meta-analysis: Addressing publication bias with model-averaging.” Psychological Method, forthcoming. doi: 10.1037/met0000405.

Mathur, M.B. & T.J. VanderWeele (2020): “Sensitivity analysis for publication bias in meta-analyses.” Journal of the Royal Statistical Society Series C 69(5): art. 10911119.

Matousek, J., Havranek, T. & Z. Irsova (2022): “Individual discount rates: A meta-analysis of experimental evidence.” Experimental Economics 25(1): pp. 318–358.

McCloskey, D.N. (2014): “Measured, Unmeasured, Mismeasured, and Unjustified Pessimism: A Review Essay of Thomas Piketty’s Capitalism in the Twenty First Century.” Erasmus Journal for Philosophy and Economics 7(2): pp. 73–115.

McShane, B.B., Bockenholt, U. & K. T. Hansen (2016): “Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes.” Perspectives on Psychological Science 11: pp. 730–749.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. & L.A. Stewart (2015): “Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.” Systematic Reviews 4(1).

Nakagawa, S., Noble, D.W.A., Senior, A.M. & M. Lagisz (2017): “Meta-evaluation of meta-analysis: Ten appraisal questions for biologists.” BMC Biology 15(1): art. 18.

Nakagawa, S., Noble, D.W.A., Lagisz, M., Spake, R., Viechtbauer, W. & A.M. Senior (2022): “A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations.” Ecology Letters 26(2): pp. 232–244.

Nakagawa, S., Yang, Y., Macartney, E.L., Spake, R. & M. Lagisz (2023): “Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences.” Environmental Evidence 12, article 8.

Neisser, C. (2021): “The Elasticity of Taxable Income: A Meta-Regression Analysis.” The Economic Journal 131(640): pp. 3365–3391.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018): “The preregistration revolution.” Proceedings of the National Academy of Sciences 115(11): pp. 2600–2606.

Oehlert, G.W. (1992): “A Note on the Delta Method.” The American Statistician 46(1): pp. 27–29.

Opatrny, M., Havranek, T., Irsova, Z. & M. Scasny (2023): “Publication Bias and Model Uncertainty in Measuring the Effect of Class Size on Achievement.” Working paper, Charles University, Prague. Available online at meta-analysis.cz/class.

Open Science Collaboration (2015): “Estimating the reproducibility of psychological science.” Science 349(6251): aac4716.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R. et al. (2021): “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 372(n71).

Pustejovsky, J.M., & E. Tipton (2022): “Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models.” Prevention Science 23: pp. 425–438.

Roodman, D., Nielsen, M.O., MacKinnon, J.G. & M.D. Webb (2019): “Fast and Wild: Bootstrap Inference in Stata Using Boottest.” The Stata Journal 19(1): pp. 4–60.

Roth, P.L., Le, H., Oh, I.-S., Van Iddekinge, C.H., & P. Bobko (2018): “Using beta coefficients to impute missing correlations in meta-analysis research: Reasons for caution.” Journal of Applied Psychology 103(6): pp. 644–658.

Rothstein, H.R. (2008): “Publication bias as a threat to the validity of meta-analytic results.” Journal of Experimental Criminology 4(1): pp. 61–81.

Rothstein, H.R., Sutton, A.J. & M. Borenstein (2005): “Publication Bias in Meta-Analysis.” In – Prevention, Assessment and Adjustments (eds. H.R., Rothstein, A.J., Sutton, M., Borenstein), John Wiley & Sons, Ltd.

Rusnak, M., Havranek, T. & R. Horvath (2013): “How to Solve the Price Puzzle? A Meta-Analysis.” Journal of Money, Credit and Banking 45(1): pp. 37–70.

Smith, G. (2018): Step Away from Stepwise. Journal of Big Data 5: p. 32.

Stanley, T.D. (2005): “Beyond publication bias.” Journal of Economic Surveys 19(3): pp. 309–345.

Stanley T.D. (2008): “Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection.” Oxford Bulletin of Economics and Statistics 70(1): pp. 103–127.

Stanley, T.D. & H. Doucouliagos (2012): “Meta-regression analysis in economics and business.” New York: Routledge.

Stanley, T.D. & H. Doucouliagos (2014): "Meta-regression approximations to reduce publication selection bias." Research Synthesis Methods 5(1): pp. 60–78.

Stanley, T.D. & H. Doucouliagos (2015): "Neither fixed nor random: Weighted least squares meta-analysis." Statistics in Medicine 34: pp. 2116–2127.

Stanley, T.D. & H. Doucouliagos (2017): "Neither fixed nor random Weighted least squares meta-regression analysis." Research Synthesis Methods 8: pp. 19–42.

Stanley, T.D. & H. Doucouliagos (2023): "Correct standard errors can bias meta-analysis." Research Synthesis Methods, https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.1631.

Stanley, T.D., Doucouliagos H., Ioannidis, J.P.A., and Carter, E. (2021). "Detecting publication selection bias through excess statistical significance." *Research Synthesis Methods*, 12: 776–795.

Stanley, T.D., Ioannidis, J.P.A. Maximilian Maier, M., Doucouliagos, H., Otte, W.M., and Bartoš F. (2023a): "Unrestricted weighted least squares represent medical research better than random effects in 67,308 Cochrane meta-analyses." Journal of Clinical Epidemiology, 157:53–58.

Stanley, T. D., Doucouliagos, H., and Havranek, T. (2023b): "Meta-analyses of partial correlations are biased: Detection and solutions," EconStor Preprints 270940, ZBW - Leibniz Information Centre for Economics.

Stanley, T.D., Jarrell, S.B. & H. Doucouliagos (2010): "Could It Be Better to Discard 90% of the Data? A Statistical Paradox." The American Statistician 64(1): pp. 70–77.

Stanley, T.D. & R.S. Rosenberger (2009): "Are Recreation Values Systematically Underestimated? Reducing Publication Selection Bias for Benefit Transfer" Bulletin of Economics and Meta-Analysis, https://www.hendrix.edu/uploadedFiles/Departments_and_Programs/Business_and_Economics/AMAES/RootnMRA.pdf.

Steel, M.F. (2020): "Model averaging and its use in economics." Journal of Economic Literature 58(3): pp. 644-719.

Strachan, R. & B. Inder (2004): "Bayesian analysis of the error correction model." Journal of Econometrics 123: pp. 307–325.

Valickova, P., Havranek, T. & R. Horvath (2015): "Financial Development and Economic Growth: A Meta-Analysis." Journal of Economic Surveys 29(3): pp. 506–526.

Vevea, J. & L.V. Hedges (1995): "A general linear model for estimating effect size in the presence of publication bias." Psychometrika 60(3): pp. 419–435.

Wetterslev, J., Thorlund, K., Brok, J. & C. Gluud (2008): "Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis." Journal of Clinical Epidemiology 61(1): pp. 64–75.

Yang, F., Havranek, T., Irsova, Z. & J. Novak (2023): "Is research on hedge fund performance published selectively? A quantitative survey." Journal of Economic Surveys, forthcoming.

Young NS, Ioannidis JPA, Al-Ubaydli O. (2008). "Why current publication practices may distort science." PLoS Medicine, 5, 134–45.

Zigraiova, D. & T. Havranek (2016): "Bank competition and financial stability: Much ado about nothing?" Journal of Economic Surveys 30(5): pp. 944–981.

Zigraiova, D., Havranek, T., Irsova, Z. & J. Novak (2021): "How puzzling is the forward premium puzzle? A Meta-Analysis." European Economic Review 134: art. 103714.

Endnotes:

[1] More than 107,000 studies published in 2022 are classified as review articles in Google Scholar and contain the word "meta-analysis". While some of them may be narrative reviews that refer to meta-analyses, many meta-analyses are not identified in this search because they are not classified as review articles. We thus consider 107,000 to represent the lower bound for the number of published meta-analyses.

[2] More than 56,000 studies published in 2022 that are classified as review articles in Google Scholar and contain the word "meta-analysis" do not contain the phrase "publication bias" or "p-hacking". Inspecting a random sample of 100 meta-analyses in more detail reveals that indeed about a half of them do not correct the data for publication bias or p-hacking.

[3] It still has value to conduct a meta-analysis if the entire literature comprises, for example, only 5 papers. But then many standard meta-analysis (and especially meta-regression) methods recommended in these guidelines cannot be used because they require a larger sample.

[4] Note that the term "fixed effects" has a different meaning in econometrics and much of the meta-analysis literature. In econometrics, a fixed-effects panel data model denotes a regression with cluster-specific dummy variables (here study-level dummies). In much of the meta-analysis literature, a fixed-effect (or common-effect) model denotes one that assumes no random heterogeneity, and outside of meta-regression typically no heterogeneity at all. The fixed-effects panel data model is more flexible than the random-effects panel data model because the latter places strong assumptions on the distribution of study-level heterogeneity terms.