Supplementary information for: "Synthesize this: Meta-analysis as a dissertation tool"

Christopher Jackson*
Andrew Q. Philips†

May 24, 2023

*christopher.jackson@colorado.edu. Graduate Student, Department of Political Science, University of Colorado Boulder, UCB 333, Boulder, CO 80309-0333.

†andrew.philips@colorado.edu. Associate Professor, Department of Political Science, University of Colorado Boulder, UCB 333, Boulder, CO 80309-0333.

Contents

1 Introduction

The preliminary stages of conducting a meta-analysis, as discussed in the manuscript, involve identifying a research question, identifying the article collection criteria, collecting articles, and coding articles. While these are arduous tasks, they are certainly some of the most important and should be taken with great care. Once the preliminary work is complete, students can move on to analyzing the data. This can differ based on the student's aims for the project and the type of data.

Below are four examples of well-constructed meta-analyses written by either graduate students working on their dissertation or early career scholars. Each example begins with a brief summary of the meta-analysis, a description of the author's collection and coding strategy, and the analysis used. While the meta-analyses presented below use some of the most common techniques used in the field, they do differ from study to study, which helps showcase how flexible this tool can be for each particular researcher's needs.

As mentioned in the manuscript, these summaries are presented only as a cursory discussion on meta-analysis and help showcase just *some* of the tools graduate students can use. By no means should this be read as an exhaustive coverage. There are many excellent in-depth guides on the subject, as detailed in Section 6. In addition, we also discuss software for those interested in carrying out a meta-analysis; see Section 7.

2 Araújo (2021)

2.1 Summary

Some argue that voters in middle- and low-income countries support political parties who implement conditional cash transfers (CCTs). Since CCTs have the potential to "alleviate poverty, increase the enrollment and attendance of kids in schools, and reduce the incidence of child mortality" (Araújo, 2021, p. 1), and they comprise a very small portion of a state's gross domestic product (ca. 0.5%), incumbents have a good reason to initiate such policies. However, some studies find contradicting results, calling into question the conclusion that CCTs always benefit incumbents on the ballot. To reconcile the divergent findings in the literature, Araújo (2021, p. 1) conducts a "meta-analysis of 10 randomized controlled trials and regression discontinuity designs (35 estimates from six countries in Latin America and Asia) to answer" whether "voters reward politicians when they implement conditional cash transfers".

2.2 Collection & Coding

Araújo (2021) describes the process of obtaining the final sample of 10 studies and 35 estimates in two phases, with different steps in each. The first phase is inclusion, which casts a wide net in order to collect as many relevant studies as possible. In the second phase, the author restricts this collected sample to account for comparability and research design quality criteria.

The first step of the inclusion phase was to start broad by using key search terms ("Conditional cash transfers or elections" and "Anti-poverty programs or elections") in the online repositories of Web of Science and Google Scholar. Araújo (2021) searched the top 30 journals in political science and economics based on Google metrics and the Scientific Journal Ranking, both from 2019. To avoid selection bias, the next step included broadening the repositories searched to include unpublished works by using the same search terms as in the first step. These repositories included the "Social Sciences Research and Network (SSRN); IDEAS/Repec; National Bureau of Economic Research (NBER); the Joint Libraries of the World Bank and the International Monetary Fund (JOLIS); and the British Library of Development Studies (BLDS)" (Araújo, 2021, p. 2). The final step of inclusion was to account for studies not written in English by applying the same search terms in the Scientific Library Online repository. The author notes that since "the high incidence of CCTs in Latin American nations, this procedure was especially important" (Araújo, 2021, p. 3).

At this point, the author had obtained 54 studies, of which 28 of these were excluded for three reasons in order to narrow the scope of the meta-analysis. First, articles that analyzed the effect of CCTs on political outcomes besides incumbent electoral support were eliminated; this was important in order to make studies "target-equivalent" (Slough & Tyson, 2022). Second, studies lacking quantitative tests of hypotheses were removed from the sample. Finally, papers that failed to include coefficients and standard

errors were disqualified. This left the author with a sample of 135 estimates from 26 studies. For these remaining articles, the author further restricted the studies by the research design used, including only those using randomized control trials (RCTs) or regression discontinuity designs (RDDs). This yielded a final sample of 10 studies and 35 estimates (see Table 1 in Araújo (2021)).

2.3 Analysis details

2.3.1 Effect size calculations

Since effect sizes are not usually reported on a standardized scale across studies, students need to convert study estimates to a comparable metric. One common approach—and used by Araújo (2021)—is to create partial correlation coefficients:[1]

$$p_{ij} = \sqrt{\frac{t_{ij}^2}{(t_{ij}^2 + \mathrm{df}_{ij})}} \tag{1}$$

where p_{ij}, the partial correlation coefficient for study i and model j, is the square root of the squared t- (or z-) statistic divided by the squared t- statistic plus the degrees of freedom from that study-model.[2] This standardizes the effect sizes between -1 and 1. Rule-of-thumb suggestions on comparing

[1] But see Godefroidt (2022) discussed below for another option.

[2] p_{ij} must be recoded as negative if the original t/z-statistic was negative.

how large effect sizes are under such standardization can be found in Cohen (2013).

The partial correlation coefficients can then be combined to create a single summary effect of the findings of a field:

$$\tilde{p} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (N_{ij} p_{ij})}{\sum_{i=1}^{I} \sum_{j=1}^{J} N_{ij}} \tag{2}$$

where p_{ij} are the partial correlation coefficients calculated in Equation 1, and N_{ij} are the assigned weights for each study-model. This creates a (weighted) average across all study-models. Weights allow graduate students to privilege the evidence provided by some partial coefficients more than others. Common weights include the number of observations (studies with more observations receive more weight, c.f., Doucouliagos & Ulubaşoğlu, 2008; Ahmadov, 2014) or the inverse of the within-study variance; others include the impact factor of the journal the article was published at or the number of citations the article has received (c.f., Philips, 2016).

Creating an overall effect size is very useful. Blair *et al.* (2021) use it to "consolidate existing evidence" on which types of commodities drive conflict, whereas Li *et al.* (2018) highlight the heterogeneity across studies of foreign direct investment based on different measures used. Effect size can even be parsed out by each study or by certain groupings; for instance Philips (2016) examines evidence of political budget cycles by the type of budgetary area

(e.g., expenditures, revenues). This is the simplest form and can be done with different types of data (e.g., continuous, binary, correlational). After Araújo (2021) excluded interactions and estimates that tested heterogeneous effects—on the basis that such coefficients cannot be directly compared across studies—the final 35 estimates from the 10 studies were transformed to partial correlation coefficients per Doucouliagos & Ulubaşoğlu (2008).

2.3.2 Common-, fixed-, and random-effects models and confidence intervals

In addition to the effect size calculation, Araújo (2021) also calculated 95% confidence intervals around the overall effect size $\tilde{p}$, which can help to highlight the dispersion of a field in disagreement (Borenstein *et al.*, 2021). However, it also requires picking an estimator. There are three main types of estimators to choose from, each of which rely on different assumptions about the effect size(s) and if/where heterogeneity is present; each of them may lead to differing confidence intervals as well:[3]

- A *Common-effect* estimator assumes there is a single, 'true' effect size. Somewhat confusingly this is also called a fixed-effect (singular, not plural) estimator. It is assumed that there is no variability between studies (other than sampling error) in terms of the effect size.
- A *Fixed-effects* (plural) estimator assumes that different studies have different effect sizes. Therefore, it is assumed that the collected studies,

[3] As discussed in the next section, these estimators are also used in meta-regression.

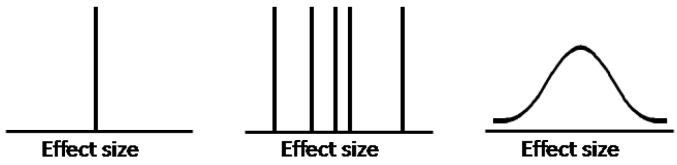
"define the entire population of interest" (Stata, 2021, p. 5) and the analyst is not seeking to generalize beyond those specific studies. In other words, "FE models are typically used whenever the analyst wants to make inferences only about the included studies" (Stata, 2021, p. 5). The effect size under a fixed-effects estimator is simply a weighted average of the individual study-specific effect sizes. Note that although the common- and fixed-effects models are "computationally identical...they differ in their target of inference and interpretation of the overall effect size" (Stata, 2021, p. 6).

- A *Random-effects* estimator assumes there is a single, 'true' effect, but that studies are sampled from this underlying population and thus exhibit some amount of heterogeneity between studies. In other words, there is variability between studies that is not assumed to exist under the common-effect model. Often this between-study variability is of substantive interest to the researcher as well. We note as well that there are different estimators that can be chosen within the random-effects framework (e.g., MLE, empirical Bayes, Hedges...); Stata's meta help documentation provides a good overview of this (Stata, 2021).

An illustration of the theoretical underpinnings of each of these three estimators is shown in Figure 1 (and originally comes from Rice *et al.* (2018, p. 209)). Borenstein *et al.* (2021, p. 77) note that since the assumption underlying a common-effect estimator "is relatively rare", most researchers instead

tend to use a random-effects model. Fixed-effects models are rare too since they do not explicitly try to generalize to a broader population in terms of the estimated effect size.[4] Indeed, Slough & Tyson (2022, p. 3) find that out of 13 recent meta-analyses in political science, 12 use a random-effects estimator while only 3 use a common-effect (they refer to it as the fixed-effect) estimator.

Araújo (2021), for example, uses the random-effects model because it "adopt[s] the assumption that there are policy implementation heterogeneity (e.g., study location or the unit selected for the intervention), and therefore calculates an average effect across studies that accounts for differences due to both chance and other factors that affect estimates" (Araújo, 2021, 3-4).



(a) Common effect model (b) Fixed effects model (c) Random-effects model

Figure 1: "Illustration of three different assumptions possibly relevant to meta-analysis" (Rice *et al.*, 2018, 209)

Araújo's (2021) results are substantively significant and indicate a causal relationship between CCTs and electoral support for the incumbent. Using common- or fixed-effects models does not change the results. His results

[4]While there are some tests for the homogeneity between studies, it has been shown to have low statistical power, so Borenstein *et al.* (2021, 78) strongly discourage students from starting with the common-effect model and then switching to the random-effects model if the test is statistically significant.

are also robust to statistical dependence whereby the author restricts the estimates included to only the highest precision per study (Araújo, 2021). Such subgroup analyses—such as parsing out measures of the "best" studies or splitting effect sizes by the type of dependent variable under analysis in a study—are very common in meta-analyses (c.f., Ahmadov, 2014; Card, 2015; Li *et al.*, 2018).

2.3.3 Meta-regression

After partial correlations have been calculated, students might also consider performing a *meta-regression analysis* (or MRA). Examples in political science range from political economy (Ahmadov, 2014; Philips, 2016; Li *et al.*, 2018) to political violence (Godefroidt, 2022) to mass/elite-public opinion (Kertzer, 2022). Since much of the variation in his meta-analysis came from differences between studies, Araújo (2021) used MRA to parse out whether study-level moderators explain the heterogeneity.

In an MRA, the dependent variable is the calculated partial correlation coefficients from Equation 1, while the independent variables are study- or model-specific features that were coded by the analyst. MRAs also involve making the random-effects vs. fixed-effects estimator choice discussed above.[5] These coded features, or "moderating variables", allow the researcher to compare how different variables, types of data, or other theoretically relevant components of a study, "cause the large variation among reported regression

[5]Or see Iršová & Havránek (2013) or Philips (2016) for an alternative Bayesian model averaging approach.

estimates" (Stanley & Doucouliagos, 2012, p. 3). For example, electoral studies scholars have used MRA to identify the different determinants of turnout rates by election type (e.g., Cancela & Geys, 2016).

A second benefit of MRA is modeling different model specifications to "directly estimate the associated misspecification biases" (Stanley & Doucouliagos, 2012, 3). Students may use different data, methods, or operationalizations of variables which may yield different results (Ahmadov, 2014; Incerti, 2020). MRA can model these differences, demonstrating how different measurements of concepts affect results. Moreover, many of these study-specific differences may be causing part of the between-study heterogeneity discussed in the previous section, and thus can help untangle whether underlying heterogeneity still exists after accounting for these factors.

A third MRA benefit is the focus on moderators. MRAs can include moderator variables not available in all the original studies; in other words, moderators that vary across but not within studies can be included. This can help avoid Simpson's paradox and is "now conventional practice among meta-analysts to include several such moderator variables" (Stanley & Doucouliagos, 2012, 89). In an MRA, a positive coefficient means that the presence of that moderator variable increases the partial correlation (all else equal), while a negative coefficient indicates that the presence of that moderator variable decreases the partial correlation. Thus, "statistical (and substantive) significance of a moderator variable suggests that it should be included in future studies...since it appears to condition the relationship between [the

two main variables of interest]" (Philips, 2016, p. 323).

In his MRA, Araújo (2021) uses four moderators, coded as dichotomous variables: RCTs (as opposed to RDDs), whether an article was peer-reviewed (relative to unpublished), whether an article was from political science (compared to economics), and whether the analysis was conducted in CCTs in Latin America (in contrast with those in Asia). He makes several conclusions, among them that, "estimates published in peer-reviewed journals tend to be smaller", that "estimates published in political science are larger than those published in economics", and "estimated effect sizes tend to be smaller in studies using RCTs" (p. 4).

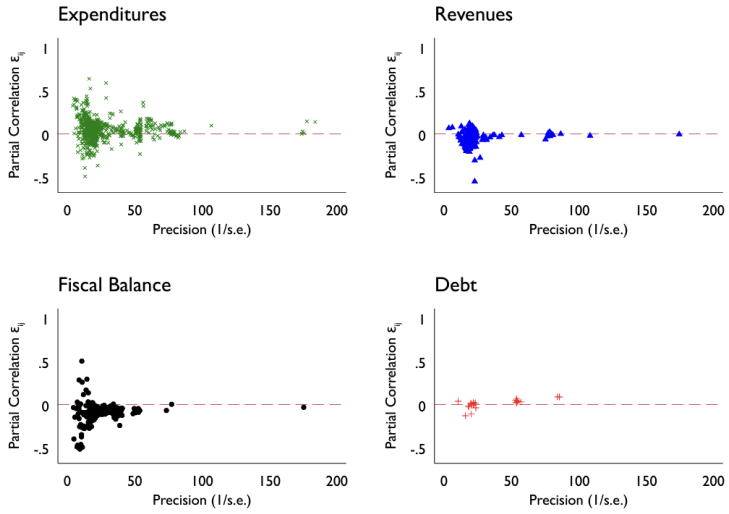
2.3.4 Publication bias

Meta-regression and effect size calculations are not the only analyses of interest. Another common focus is on publication selection bias, which can arise if estimates are reported (and manuscripts ultimately published) based off the statistical significance of the effects alone. If it exists, "valid empirical inference is threatened because the research base will not be a representative sample of the population of estimates for the phenomenon in question. As a result, all conventional summaries will be biased" (Stanley *et al.*, 2010, p. 70). Meta-analyses are powerful since although most scholars know about publication or 'file drawer' biases, this approach allows analysts to calculate the extent to which it may actually exist in a body of literature. Moreover, some approaches even allow the analyst to correct for such bias.

Funnel-asymmetry plots/tests use effect size calculations to ascertain whether publication bias exists using a *funnel-asymmetry test* (FAT) (Stanley & Doucouliagos, 2012, 62-68). The first step is to create a funnel plot, an example of which is shown in Figure 2. Here, the partial correlation coefficients are plotted on the vertical axis, while the precision of these estimates—one divided by the standard error of the estimate—is plotted on the horizontal axis.[6] Standard errors are a function of the partial correlation as well as the degrees of freedom for study-i, model-j: $\text{SE}_{ij} = \sqrt{\frac{1-p_{ij}}{df_{ij}}}$. Thus, the intuition behind the funnel plot is that smaller-N studies with larger standard errors (lower precision) should appear on the left-hand side of the plot while larger-N studies with smaller standard errors (high precision) should be on the right-hand side. Bias, if it exists, is typically associated with a non-funnel shape, especially for the low-precision studies. If there are found to be "missing" (i.e., file-drawer or publication bias) studies that, if published, would make the funnel plot symmetrical (Borenstein *et al.*, 2021), a non-parametric *trim-and-fill* method can be used which estimates the number of missing studies, imputes them, and recalculates the overall effect size (Duval & Tweedie, 2000a); we discuss this more in another example below.

A corresponding statistical test—the FAT—involves effectively a weighted regression of the funnel plot (c.f., Stanley & Doucouliagos, 2012). This involves a linear regression of the effect size regressed on its standard error

[6] These are often plotted inversely; i.e., the precision is on the vertical axis and partial correlations on the horizontal.



Notes: 1198 total estimates for 88 studies.

Figure 2: Funnel plots of the four budgetary categories from Philips (2016)

weighted by the inverse variance:

$$\frac{\hat{\theta}_k}{SE_{\hat{\theta}_k}} = \beta_0 + \beta_1 \frac{1}{SE_{\hat{\theta}_k}} \tag{3}$$

where $\hat{\theta}_k$ are the observed effect sizes over their respective standard error. This is regressed on by the precision, or the inverse of the standard error. For this test, the z-score should the distributed around zero. If the test reveals an overdispersion of low-precision studies with statistically significant z-scores, then publication bias is possible. For an in-depth discussion see Rothstein *et al.* (2005), Higgins *et al.* (2019) or Vevea *et al.* (2019).

Araújo (2021) accounts for publication selection bias with plots and FAT.

He finds little evidence of publication bias, suggesting that the literature has correctly concluded that voters reward politicians for the implementation of CCTs.

3 DeCrescenzo (2020)

3.1 Summary

As a whole, DeCrescenzo's (2020) dissertation examines the conventional wisdom in American Politics that political candidates should mimic the policy preferences of their district's median voter. He argues that primary voters tend to hold more extreme values than the median electorate, and therefore candidates will adopt these positions. The dilemma candidates face is balancing appearing too moderate in the primary with being too partisan in the general election. Using Bayesian methods, DeCrescenzo (2020) finds that candidates lean into partisan policy preferences during primary elections while voters tend to prefer candidates who represent the "ideological core" of the party.

While DeCrescenzo (2020) does not conduct an original Bayesian meta-analysis as a stand-alone chapter or section, he does use the method to build on the findings of a previous meta-analysis from Green *et al.* (2016); these authors present four different experiments in a single study, which they synthesize using a meta-analysis.

3.2 Collection & Coding

Since his is not an original meta-analysis, the author did not need to collect or code studies. Instead, DeCrescenzo (2020) used the data from Green *et al.*'s (2016) fixed-effects meta-analysis. Their study looks at the use of yard signs in campaigns of various salience and setting: "both primary and general election campaigns, races as high-profile as governor and as low-profile as county commissioner, electorates in different states, and more" (DeCrescenzo, 2020, p. 117-118). Green *et al.* (2016, p. 148) used a fixed-effects estimator since they were interested in the "precision-weighted average of the four estimated direct treatment effects," although keep in mind this estimator limits generalizability to other studies in what could be considered the broader population.[7]

3.3 Analysis details

DeCrescenzo's (2020) analysis critiques the meta-analysis performed by Green *et al.* (2016) because he argues that the fixed-effects model is inappropriate due to the assumptions behind the fixed-effects model, and that there might be between-study heterogeneity of substantive interest. To improve on this, he uses a Bayesian model because it "exposes this prior [no cross-study heterogeneity which is highly specific], allowing the researcher to scrutinize and improve [Green *et al.*'s (2016)] model" (DeCrescenzo, 2020, 122).

[7] This is discussed more in the Section, "Common-, fixed-, and random-effects models and confidence intervals".

What distinguishes DeCrescenzo's (2020) meta-analysis from Green *et al.*'s (2016) is the use of priors, specifically three that do not "secretly posit a highly specific prior by assuming no cross-study heterogeneity" (DeCrescenzo, 2020, 122). Instead of using "naïve" flat or ignorant priors, DeCrescenzo (2020, 119) uses three different priors with varying levels of μ and σ, where σ is always greater the μ to allow for cross-study heterogeneity:[8]

$$\text{Agnostic}: \quad \mu \sim \text{Normal}(0, 0.05) \tag{4}$$

$$\text{Skeptical}: \quad \mu \sim \text{Normal}(0, 0.01) \tag{5}$$

$$\text{Optimistic}: \quad \mu \sim \text{Normal}(0.05, 0.05) \tag{6}$$

While some may charge that these priors treat the data unfairly, other currently used and popular priors (e.g., Cauchy or Student T's) would have resulted in greater uncertainty with respect to the population treatment effect (DeCrescenzo, 2020). Moreover, the ignorant prior would have resulted in more uncertainty in the posterior distribution than the three priors used. Since "many of these priors result in population estimates that fail to reject the null hypothesis... the Bayesian model exposes this prior, and allows the researcher to scrutinize and improve their model" (DeCrescenzo, 2020, 120, 122). The results indicate that *only* the fixed-effects model, and the implicit priors behind it (i.e., that of Green *et al.* (2016)), rejects the null hypothesis;

[8] Contrast this with Green *et al.*'s (2016) "highly restrictive" prior that limits σ to 0.

the other more relaxed priors result in a failure to reject the null hypothesis.

4 Godefroidt (2022)

4.1 Summary

A large body of literature has explored the effects terror attacks have on individual political attitudes, especially after September 11, 2001. Despite the many studies on the topic, it is not clear how terrorism affects political attitudes. Extant work has produced varying effect sizes. On the one hand, some contend that attitudes post-terror attacks can jeopardize democratic stability, on the other hand, some work finds that terror attacks have only acute effects on citizens' attitudes. Moreover, even if these discrepancies can be resolved, the generalizability of the body of work is unclear. As there is no systematic review of this literature, Godefroidt (2022) conducts the first meta-analysis on the effect of terrorism on social and political attitudes using an impressive sample discussed below.

4.2 Collection & Coding

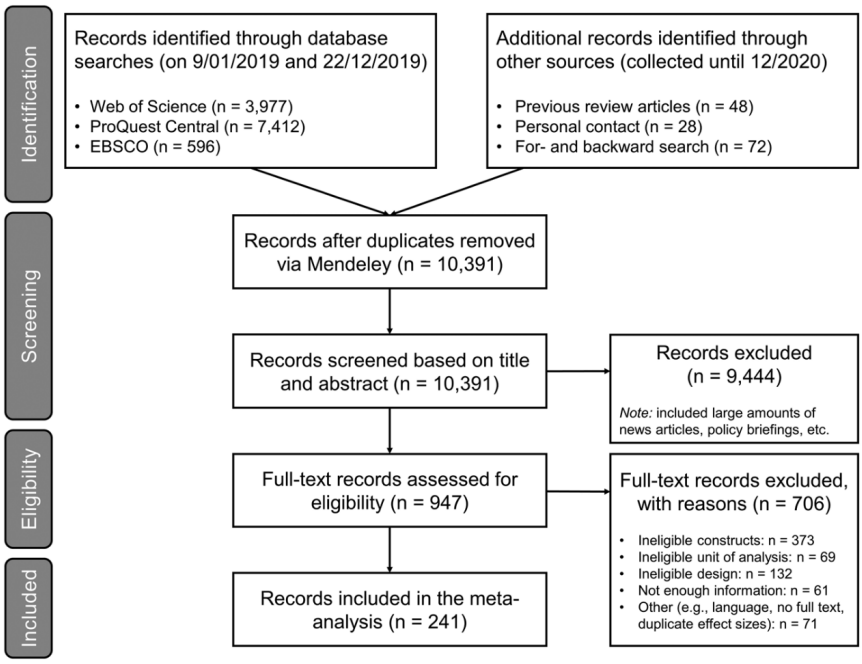
Godefroidt (2022) used a four-fold search strategy. First, she used an extensive search term: "(prejudice OR stereotyp* OR out-group OR attitud* OR authoritarian* OR conservat* OR"public opinion" OR "policy support" OR "political consequences" OR "political tolerance" OR ideolog* OR voting

OR vote*) AND (terror* OR attack* OR "political violence" OR bomb* OR "September 11" OR "9/11" OR"March 11" OR "Charlie Hebdo" OR "Paris attacks" OR "Utoya" OR "Utøya")" (SI Appendix §B.1, pp. 3). This term was entered in three databases: Web of Science, ProQuest, and EBSCO. Within these repositories, databases included disciplines such as political science, sociology, and criminology.

Second, Godefroidt (2022) sent out for calls for working or unpublished papers on Twitter, to listservs, relevant societies (e.g., American Political Science Association, European Political Science Association, and Society of Terrorism Research, etc.), and to "33 prominent scholars in the field" in order to broaden her search. Third, she also looked through four articles that were qualitative reviews of the literature (i.e., those that did not conduct a meta-analysis). Fourth, Godefroidt (2022) did a forward and backward search on about half of the articles' reference lists. These four steps yielded a total of 12,133 articles, and 10,391 after deleting duplicates. Less than 1/10 (947) of these articles were retained, as the remainder did not consist of a quantitative component. Using five inclusion/exclusion criteria, construct, units, study, design, and statistics (SI Appendix §B.1, pp. 4), Godefroidt (2022) produces a final sample of 241 articles that estimated a total of 1,733 effect sizes.[9] An illustration of this process from Godefroidt (2022, p. 5) is shown in Figure 3.

[9]Godefroidt (2022) notes a total of 326 studies, since more than one study is possible within a given manuscript.

FIGURE 1 PRISMA Flowchart of Selection Process



Note: The flowchart shows the meta-analysis data-collection process. "Records excluded" were excluded because the title or abstract did not reflect the subject matter of the meta-analysis or because the records appeared to be incomplete, unavailable, or nonacademic. "Full-text articles excluded" were excluded due to a failure to meet the inclusion criteria (see SI Table B.1, p. 4). The "Records included in the meta-analysis" refers to a published or unpublished collection of unique studies.

Figure 3: Collection criteria flowchart from Godefroidt (2022, p. 5)

4.3 Analysis details

As is often the case when conducting a meta-analysis, the studies report different associations between key explanatory variables and dependent variables. These tend to include correlation coefficients, regression coefficients, odds ratios, mean differences, etc. (Godefroidt, 2022, 4). Consequently, Godefroidt (2022) converts these to Pearson's correlation coefficients. Next, correlation coefficients were given appropriate signs to correspond with positive or negative relationships between terrorism and political attitudes. Finally, given the non-normal sampling distribution, the author used a Fisher's Z transformation using Equation 7 (Godefroidt, 2022, 4):

$$z = \frac{1}{2} \log \frac{1+r}{1-r} \tag{7}$$

Like partial correlation coefficients (c.f., the example from Araújo, 2021), Z-scores are one form of creating comparable metrics of effects across studies.

4.3.1 Model Specification and Analysis

Godefroidt (2022) used a random-effects, three-level meta-analysis. In contrast to some of the examples discussed above, which parse heterogeneity into two-levels, a three-level meta-analysis has three assumptions: "that observed effect sizes differ because of (1) sampling variance, (2) variance between manuscripts, and (3) variance between the correlations from within

the same manuscript" (Godefroidt, 2022, 4). This type of analysis allowed the author to model heterogeneity both within and between studies in addition to including different moderator variables. With it, Godefroidt (2022) calculates overall effect sizes that suggest that terrorism is most linked to outgroup hostility and conservative shifts, and to a lesser extent terrorism is linked to rally effects (p. 8); she notes that while these effects are small substantively speaking, they are statistically significantly different from no effect.

Godefroidt (2022) also uses a meta-regression analysis (which was first discussed in the Araújo (2021) example). She includes several moderators of interest for those studying terrorism and political attitudes, including ideological (e.g., Islamist), methodological/research design (e.g., experiment, observational data, student survey, convenience sample), and country of analysis (e.g., US, Israel). She finds several interesting sources of heterogeneity, primarily in ideology and research-design based decisions, especially for those looking at outgroup hostility and conservative shifts (see Godefroidt (2022), Figure 3). Godefroidt (2022) also plots effect sizes over publication year, and finds that the effect sizes in the literature appear to attenuate over time.

4.3.2 Publication Bias

While only briefly mentioned in the main manuscript, Godefroidt (2022) also used several diagnostics for publication bias using her meta-analysis data.

First, since Godefroidt (2022) collected unpublished studies, she could use

MRA to examine if this led to statistically significant differences (in terms of effect sizes) from published studies. She finds almost no evidence of such publication selection bias using a MRA.

Second, she creates funnel plots—these are shown in Figure 4. Visually, most of the effect sizes appear to be symmetric, which suggests little evidence of publication bias. However, Godefroidt (2022) also uses the trim-and-fill method, which is a further statistical test for publication bias (Duval & Tweedie, 2000a,b). This approach "estimates what studies might be missing and then adds them to the analysis" (Borenstein *et al.*, 2021, 321). Figure 4 shows three funnel plots: the grey points are from the funnel plot, the black points are the estimated points from the trim-and-fill method, which mirrors the right-hand side of Plot C. Applying trim-and-fill to Figure 2, the result of this method then may alter the overall effect size, which can be recalculated after imputing these 'missing' studies.

Third, although she does not find much evidence of publication bias, Godefroidt (2022) employs a tool to both test for and correct for publication bias. After testing for publication bias, researchers should also investigate whether there is an empirical effect beyond any potential effect caused by publication bias. *Precision-effect testing* (PET, sometimes called FAT-PET since it is typically estimated in conjunction with the funnel asymmetry test) is a form of MRA where the standard error acts as a moderator variable and the intercepts are the estimate when the standard error is zero. (A standard error of zero would, in theory, represent an infinite sample size.) Similar to

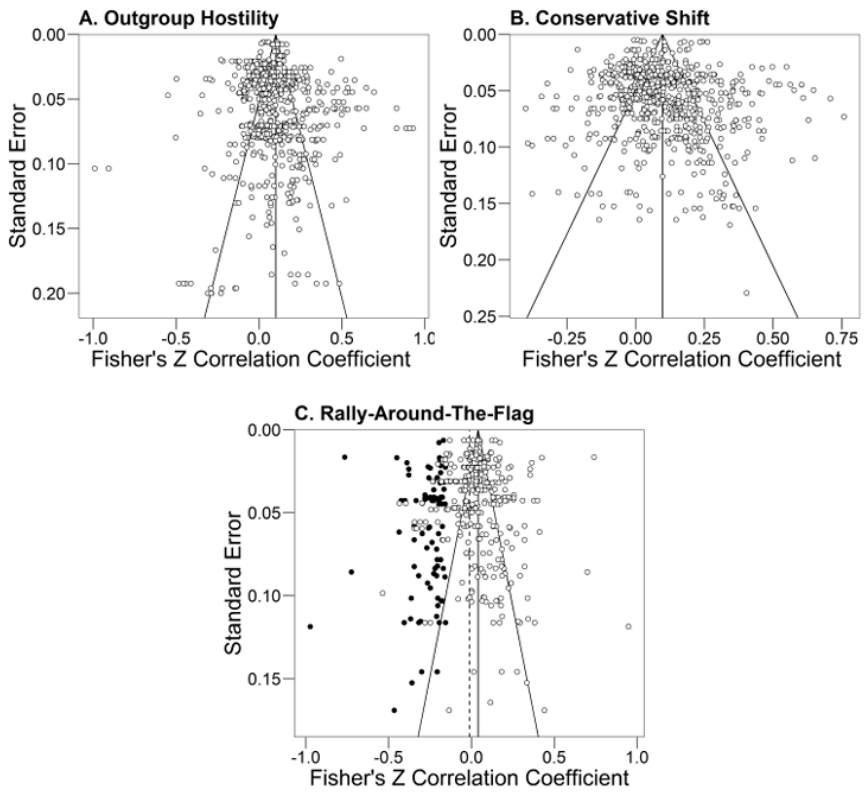


Figure 4: Trim-and-fill method applied to funnel plots (Godefroidt, 2022, SI Figure C.1, p. 17)

this is the *precision-effect estimate with standard error*, or PEESE model, which provides an effect size estimate when the variance is zero. While the standard errors and variances are included in the PET and PEESE models, respectively, the coefficient of interest is the intercept.[10] Formally, PET regresses the t-statistic of a study-model effect on its respective standard error (Stanley & Doucouliagos, 2012; Alinaghi & Reed, 2018):

$$t_{ij} = \beta_0 + \beta_1 SE_{ij} + \epsilon_{ij} \tag{8}$$

Equation 8 is estimated using weighted least squares (WLS, where weights are $\frac{1}{SE_{ij}}$), the weighting being done so as to make the error variance homoskedastic. In other words, estimate:

$$\frac{t_{ij}}{SE_{ij}} = \beta_0 \frac{1}{SE_{ij}} + \beta_1 + \frac{\epsilon_{ij}}{SE_{ij}} \tag{9}$$

The FAT can be obtained by testing whether $\hat{\beta}_1 = 0$, while the PET is obtained by testing whether $\hat{\beta}_0 = 0$. Due to WLS $\hat{\beta}_0$ can be seen as the coefficient of precision (recall precision was $\frac{1}{SE_{ij}}$). For PET, if "the intercepts are of a similar magnitude and significance of the overall effect sizes, the

[10] Although Borenstein *et al.* (2021, 327) caution that "the reader should be skeptical of several increasingly popular methods for examining publication bias known as *p-curve* and PET-PEESE which maintain that the publication-bias adjusted effect is the true effect" because "one should never assert that adjusted value is the 'Correct' value" (Borenstein *et al.*, 2021, 327), Stanley & Doucouliagos (2012, 61) note that the "estimates of β_0 from both equation [9] and [10] have been shown to be among the best in comprehensive simulations of alternative corrections for publication bias".

results prove to be robust" (Godefroidt, 2022, SI, p. 18).

The PEESE model is similar to Equation 9, but uses the variance instead of the standard error, which (after the same WLS transformation) is given as (Stanley & Doucouliagos, 2012; Alinaghi & Reed, 2018):

$$\frac{t_{ij}}{SE_{ij}} = \beta_0 \frac{1}{SE_{ij}} + \beta_1 SE_{ij} + \frac{\epsilon_{ij}}{SE_{ij}} \tag{10}$$

PEESE is an improvement on PET in that it provide a better estimate of the underlying effect in the presence of publication bias (c.f., Stanley *et al.*, 2007; Stanley & Doucouliagos, 2012, 2014). As with PET, the test of interest under PEESE is whether $\hat{\beta}_0 = 0$. Using these tests, Godefroidt (2022) finds limited evidence of publication selection bias in the literature ("with an exception for the rally-'round-the-flag subsample" (SI Table C.9, p. 17 and Table C.10, p. 18)), and still finds evidence of a true underlying effect even after trying to correct for any publication bias.

Last, Godefroidt (2022) also runs several other robustness checks detailed in the online appendix. First, outliers did not affect the results (SI Table C.6, p. 14). Second, the results are robust to different model specifications (SI Table C.5, p. 13). Third, Godefroidt (2022) checks study quality (SI Table C.7, p. 15). Fourth, correlations that came from regression coefficients were excluded (SI Table C.8, p. 16). Overall, the results suggest that there is some evidence that terrorism affects political attitudes. Terrorist attacks are found to increase outgroup hostility and political conservatism while only mildly

increasing a rally-'round-the-flag effect. A key note of caution that Godefroidt (2022) warns about is the type of terrorism and the society in which attitudes are measured matters; much of the research has been conducted on *Islamist* terrorism and responses in *Western* societies.

5 Incerti (2020)

5.1 Summary

Given the divergent findings across experimental studies examining whether voters punish politicians for corruption at the ballot, Incerti (2020) conducts a meta-analysis to reconcile these discrepancies. The key argument is that the varying results stem primarily from study design and methodological differences. Whereas survey experiments may inflate results due to social desirability bias and the lower perception of costs, field experiments may underestimate results as a consequence of "weak treatments and noncompliance" (Incerti, 2020, p. 761).

5.2 Collection & Coding

Incerti (2020) used the search terms ('"corruption experiment," "corruption field experiment," "corruption survey experiment," "corruption factorial," "corruption candidate choice," "corruption conjoint," "corruption, vote, experiment," and "corruption vignette"') in databases and by following citation

chains to identify articles. While he did not restrict the search by discipline, only papers from political science and economics made it into the final sample. To avoid publication selection bias, Incerti (2020) includes both published and working papers. There was no mention of the selection process for non-English studies, though there appear to be studies included that were conducted in different countries.

Beyond these search criteria, studies were included or excluded as they fit with the research question. Per the author's online appendix (A.2), 10 studies were excluded for the following five reasons: "Lack of no-corruption control group" (6), "Outcome is hypothetically changing actual vote" (1), "Outcome is favorability rating, not vote share" (1), "Data identical to Weitz-Shapiro and Winters (2013)" (1), "Data identical to Weitz-Shapiro and Winters (2017)" (1). For example, papers that examine electoral fraud are excluded, as the debate between whether this is considered either clientelism or corruption is unresolved. Another example is the inclusion of a study that looks at "politicians' asset accumulation and criminality, which may imply corruption but is not as direct as other types of information provision" (Incerti, 2020, 763).[11] The author lists several other cases that were included for similar reasons, as well as how excluding theses studies affects the results (see the online appendix). In total, "10 field experiments from 8 papers, and 18 survey experiments from 15 papers" were included in the meta-analysis (Incerti, 2020, 761).

[11] Although the results do not change when excluded.

5.3 Analysis details

Using field and survey experiments, Incerti (2020) conducted both fixed-effects and random-effects meta-analyses. The results suggest that field experiments produce null findings whereas survey experiments are strongly negative indicating that voters punish corrupt politicians. Testing for heterogeneity, he includes a dummy variable for type of experiment which suggests that about 68% of the variation in results can be explained by study type.

5.3.1 Publication Bias

Incerti (2020) used several tests for publication bias, which we have described above in previous sections. These include a funnel plot and FAT test, PET-PEESE tests, and trim-and-fill methods (for funnel plots). While he finds some evidence for asymmetry when examining funnel plots, much of this "asymmetry disappears when accounting for heterogeneity by type of experiment" (Incerti, 2020, p. 766). Nor do PET-PEESE estimates provide much evidence of publication selection bias.

Incerti (2020) also examines p-curves, which is another method used to test for publication selection bias. The procedure involves plotting the distribution of p-values from included studies. Distributions centered around a p-value of 0.05 suggest evidence of p-hacking (Veling *et al.*, 2020). If the distribution is skewed to the right near 0.01 then it is closer to the true effect (Incerti, 2020; Simonsohn *et al.*, 2014). The p-curve indicates no publication

bias for survey experiments, although is not viable for field experiments due to the small sample size.

After examining four different approaches to publication selection bias, the author contends that there is still not conclusive evidence that publication bias is *not* a factor, especially given the small sample. Nonetheless, the results suggest that the heterogeneity is likely due to study design instead such as social desirability and hypothetical biases, noncompliance, and the salience and strength of treatment. Since these are not tested and pertain more to experimental design than to meta-analysis, these elements of Incerti's study—which he examines in further detail in his article—are not discussed here.

6 Annotated/suggested reading list

1. Lipsey & Wilson (2001): This book offers a great start for students who want to use meta-analysis. The book explains how to identify the correct exclusion and inclusion criteria based on the theoretical underpinnings of the meta-analysis, how to code the final sample of studies, when to use a meta-analysis, and how to analyze and interpret the meta-analysis using software.

2. Stanley & Doucouliagos (2012): *Meta-Regression Analysis in Economics and Business* is an introductory text for students new to meta-analysis. The text summarizes some best-practices to set new practitioners on

the right path. It explains what meta-regression is, when to use it, and the different nuances of the tool (for example, tests for publication selection bias).

3. Borenstein *et al.* (2021): The second edition of *Introduction to Meta-Analysis* provides students with a thorough examination of meta-analysis, and is widely considered *the* meta-analysis textbook. It details how to transform study effect sizes for synthesis, the differences between fixed-effect and random-effects models, and how to understand heterogeneity. The new edition also provides readers with guides on how-to as well as software recommendations.

4. Slough & Tyson (2022): This articles aims to provide a clear explanation for the role of external validity for meta-analysis. The external validity of individual studies is often unclear, thus researchers turn to meta-analysis. But there is some debate about whether these are always generalizable. The authors contend that "literal" equality between studies is less important than harmonization among construct and measurement because the latter must be represented across all studies in the meta-analysis sample. These two harmonizations are considered as more important than others because the student can change these via "design or inclusion criteria".

7 Software for conducting meta-analysis

There are several options for software for students to use when conducting a meta-analysis. We list four below.

7.1 R

Students who wish to use R can install Wolfang Viechtbauer's pacakge, `Metafor`. "The package consists of a collection of functions that allow the user to calculate various effect size or outcome measures, fit equal-, fixed-, random-, and mixed-effects models to such data, carry out moderator and meta-regression analyses, and create various types of meta-analytical plots" and can be downloaded at https://www.metafor-project.org/. The program also includes code to replicate the examples from the first edition of *Introduction to Meta-Analysis*, available at: https://wviechtb.github.io/metaanalysisbooks/borenstein2009.html

7.2 Stata

In Stata, students can use the commands already in the program; the 17th edition of Stata contains commands such as meta summarize (to combine studies and compute overall effect sizes), meta forestplot (to create forestplots of effect sizes across studies), meta funnelplot (to create funnelplots), and meta mvregress (for metaregression analysis). These are accessible from the

Statistics dropdown menu. To learn about commands for meta-anlysis in Stata, visit: https://www.stata.com/features/meta-analysis/.

Borenstein *et al.* (2021) also note that older user-written commands exist as well. They recommend the second edition of *Meta-Analysis in Stata: An Updated Collection from the Stata Journal* (Sterne, 2009) for a guide to these commands.

7.3 Other resources

Biostat, Inc. has created its own software, Comprehensive Meta-Analysis (CMA). It was initially released in 2000 and is now in version 4. For information on this program, visit www.Meta-Analysis.com. While this program provides access to free trials and lecture and worked example videos to accompany the trial, it is not free to access.

The last example is `Revman` which is used and produced by the Cochrane Collaboration. This is web-based software that can be tried for free for a month. Students only need to make an account to try the limited version. `Revman` can be accessed here: https://training.cochrane.org/online-learning/core-software/revman

References

Ahmadov, Anar K. 2014. Oil, democracy, and context: A meta-analysis. *Comparative Political Studies*, **47**(9), 1238–1267.

Alinaghi, Nazila, & Reed, W Robert. 2018. Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research synthesis methods*, **9**(2), 285–311.

Araújo, Victor. 2021. Do Anti-Poverty Policies Sway Voters? Evidence from a Meta-Analysis of Conditional Cash Transfers. *Research & Politics*, **8**(1).

Blair, Graeme, Christensen, Darin, & Rudkin, Aaron. 2021. Do commodity price shocks cause armed conflict? A meta-analysis of natural experiments. *American Political Science Review*, **115**(2), 709–716.

Borenstein, Michael, Hedges, Larry V., Higgins, Julian P. T., & Rothstein, Hannah R. 2021. *Introduction to Meta-Analysis*. John Wiley & Sons.

Cancela, João, & Geys, Benny. 2016. Explaining Voter Turnout: A Meta-Analysis of National and Subnational Elections. *Electoral Studies*, **42**(June), 264–275.

Card, Noel A. 2015. *Applied meta-analysis for social science research*. Guilford Publications.

Cohen, Jacob. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

DeCrescenzo, Michael G. 2020. *Do Primaries Work? Constituent Ideology and Congressional Nominations*. The University of Wisconsin-Madison.

Doucouliagos, Hristos, & Ulubaşoğlu, Mehmet Ali. 2008. Democracy and economic growth: a meta-analysis. *American journal of political science*, **52**(1), 61–83.

Duval, Sue, & Tweedie, Richard. 2000a. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the american statistical association*, **95**(449), 89–98.

Duval, Sue, & Tweedie, Richard. 2000b. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**(2), 455–463.

Godefroidt, Amélie. 2022. How Terrorism Does (and Does Not) Affect Citizens' Political Attitudes: A Meta-Analysis. *American Journal of Political Science*.

Green, Donald P., Krasno, Jonathan S., Coppock, Alexander, Farrer, Benjamin D., Lenoir, Brandon, & Zingher, Joshua N. 2016. The effects of lawn signs on vote outcomes: Results from four randomized field experiments. *Electoral Studies*, **41**(Mar.), 143–150.

Higgins, Julian PT, Thomas, James, Chandler, Jacqueline, Cumpston, Miranda, Li, Tianjing, Page, Matthew J, & Welch, Vivian A. 2019. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Incerti, Trevor. 2020. Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design. *American Political Science Review*, **114**(3), 761–774.

Iršová, Zuzana, & Havránek, Tomáš. 2013. Determinants of horizontal spillovers from FDI: Evidence from a large meta-analysis. *World Development*, **42**, 1–15.

Kertzer, Joshua D. 2022. Re-assessing elite-public gaps in political behavior. *American Journal of Political Science*, **66**(3), 539–553.

Li, Quan, Owen, Erica, & Mitchell, Austin. 2018. Why do democracies attract more or less foreign direct investment? A metaregression analysis. *International Studies Quarterly*, **62**(3), 494–504.

Lipsey, Mark W, & Wilson, David B. 2001. *Practical meta-analysis.* SAGE publications, Inc.

Philips, Andrew Q. 2016. Seeing the Forest Through the Trees: A Meta-Analysis of Political Budget Cycles. *Public Choice*, **168**(3), 313–341.

Rice, Kenneth, Higgins, Julian PT, & Lumley, Thomas. 2018. A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**(1), 205–227.

Rothstein, Hannah R, Sutton, Alexander J, & Borenstein, Michael. 2005. Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 1–7.

Simonsohn, Uri, Nelson, Leif D, & Simmons, Joseph P. 2014. P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, **143**(2), 534.

Slough, Tara, & Tyson, Scott A. 2022. External Validity and Meta-Analysis. *American Journal of Political Science*.

Stanley, T. D., & Doucouliagos, Hristos. 2012. *Meta-regression Analysis in Economics and Business*. Routledge.

Stanley, TD, Doucouliagos, Hristos, *et al.* 2007. Identifying and correcting publication selection bias in the efficiency-wage literature: Heckman meta-regression. *Economics Series*, **11**, 2007.

Stanley, Tom D, & Doucouliagos, Hristos. 2014. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, **5**(1), 60–78.

Stanley, Tom D, Jarrell, Stephen B, & Doucouliagos, Hristos. 2010. Could it be Better to Discard 90% of the Data? A Statistical Paradox. *The American Statistician*, **64**(1), 70–77.

Stata. 2021. *Stata 17 Meta Reference Manual*. StataCorp. College Station, TX: Stata Press.

Sterne, Jonathan AC. 2009. *Meta-analysis in Stata: an updated collection from the Stata Journal*. StataCorp LP.

Veling, Harm, Chen, Zhang, Liu, Huaiyu, Quandt, Julian, & Holland, Rob W. 2020. Updating the p-curve analysis of Carbine and Larson with results from preregistered experiments. *Health Psychology Review*, **14**(2), 215–219.

Vevea, Jack L, Coburn, Kathleen, & Sutton, Alexander. 2019. Publication bias. *The handbook of research synthesis and meta-analysis*, **3**, 383–432.