**META FAQs**

**How to write well**

You will make your living by writing in one form or another. The more you read, the better your writing gets. One of the best writers I have read is Erik Larson (In the garden of the beasts, Isaac’s storm, …). But that’s a life-long process of improvement. My first drafts are horrible (see this document), writing gets better in revisions. If you want to have a well-written thesis, you will have to revise it constantly, restructuring sentences, paragraphs, sections. Watch this important contribution to writing by prof. Pinker (Harvard) at <https://www.youtube.com/watch?v=OV5J6BfToSw>. The bottom-line: write in simple terms and know your audience. Language evolves constantly. Books, especially those on writing, get obsolete. If in doubt, “google” the expression. Check what most people use right now: that's how language gets formed.

As non-native speakers, we need help when we decide to write in English but the following books will be handy even for people writing in Czech: everybody should read the book of Strunk & White (2000): “The Elements of Style” (available in audiobook as well). McCloskey (1999): “Economical Writing” is recommended for all economists; you can read it in one afternoon. For those aiming further in their academic career or aiming for prices and distinctions, there is a book by Williams (2010): “Style: Lessons in Clarity and Grace.” Find older publicly available book versions here:

* <https://www.johnhcochrane.com/research-all/writing-tips-for-phd-studentsnbsp>,
* <https://www.dropbox.com/s/ay77u6ebgm51ogh/reading_strunkwhite.pdf?dl=0>,
* <https://www.dropbox.com/s/q1eal6f14jh0o3w/reading_mccloskey.pdf?dl=0>,
* <https://www.dropbox.com/s/0r3ugtomikg9qi5/reading_mccloskey2.pdf?dl=0>,
* <https://www.dropbox.com/s/91o69t9vj7bvwin/reading_williams.pdf?dl=0>.

Get inspired by well-written theses of your former colleagues:

* https://dspace.cuni.cz/handle/20.500.11956/124595
* https://dspace.cuni.cz/handle/20.500.11956/94855

You will write **1) introduction, 2) conclusion, and 3) abstract** AFTER you finish everything else in your thesis. These parts are by far the most important ones and have to be very well written. Read our latest papers here:

* http://meta-analysis.cz/students/
* http://meta-analysis.cz/risk/
* http://meta-analysis.cz/skill/

Write in English and in LaTeX. You can download the latest template here:

<https://is.cuni.cz/studium/predmety/index.php?do=download&did=245904&kod=JEM213>

**To correct your grammar, use ChatGPT at chat.openai.com. Please!** The technology is still not good enough to produce an academic text (YET :)) but it is fantastic for rephrasing your text, helping to create summaries, discussions, and so on. It can even help you with codes for statistical packages such as R and Stata. It can identify problems in LaTeX code. It is free of charge so use it to your advantage.

**What is the basic literature for how to conduct a meta-analysis?**

Read the cookbook and this FAQ. Read our latest meta-analyses (where data and codes are provided, you can try first to replicate one of them, an easy one is meta-analysis.cz/students/). If you have further questions, you can read [Borenstein´s](https://www.agropustaka.id/wp-content/uploads/2020/04/agropustaka.id_buku_Introduction-to-Meta-Analysis.pdf) or [Stanley](https://www.dropbox.com/s/u87u9flooj6odz7/Meta-Regression-Analysis-in-Economics-and-Business-by-T.D.-Stanley-Hristos-Doucouliagos-.pdf?dl=0)´s textbooks on meta-analysis but the research is going forward very fast. There are many free sources, of course:

<https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/>, <https://www.maer-net.org/blog>.

**I have heard about publication bias and p-hacking. Is it the same thing? Is it possible to differentiate between these two in your analysis?**

Conceptually, publication bias and p-hacking are distinct terms. P-hacking denotes researchers’ effort to produce statistically significant results, and often stems from publication bias. Thus, the term **p-hacking** is usually reserved for original researchers and their specification search (meaning: this model does not give me the statistically significant results I need for a particular variable so let’s try another model that will give me the significance level I need for a particular variable) or adding observations to their dataset one by one to get to statistical significance. **Publication bias** as such usually refers to “the whole package of trouble” on the level of the original researchers/authors (with motivation: I want to write a story that is publishable in a journal) as well as editors/publishers of journals (with motivation: I want to publish a story that would be well cited). Publication bias is this idea that some estimates have higher probability of getting reported than other estimates. The preference always stems from looking for better story (and it does not matter who wants better story, authors or editors); meaning, it is better to have statistically significant estimates than statistically insignificant, or it is better to have estimates that are in line with some strong theory and do not look suspicious (the strongest theory in economics would be the law of demand, meaning you should not get a positive estimate for the price elasticity of demand for some good that is clearly an ordinary one and not Giffen, eg. milk). Please note that publication bias per se does not necessarily mean people involved in the publication process cheat. A researcher may have a great reason to publish only their best performing model (and leave the rest for the appendix). So for a single study it would make perfect sense to pick and choose. The trouble starts when “picking and choosing” is systematic across the literature and then, it is a job for the meta-analyst to report well on the literature.

It is actually unfeasible in empirical work to separate these two effects (you would have to compare the articles at their earliest stages of writing with the articles at their submission stage being submitted to journal with the articles that were actually published). These effects tend to be, in the end, observationally equivalent. Applied meta-analysts thus typically use the term publication bias more generally to also include p-hacking, and you should follow this practice.

Note that sometimes people mismatch publication bias for this idea that unpublished results (from working papers, books, basically those documents that are not peer-reviewed) do not suffer from publication bias but published results (those published in journals that are peer-reviewed) do suffer from publication bias. This is untrue. It could be the case but rarely is. That is why we prefer to call the phenomena rather “publication selection bias” because it is the “selection” part that matters not the “publication” part. Selection can already start at the level of an unpublished study (which precedes publication in a journal), so it starts at the level of the authors, and can proceed unnoticed to the level of editors. In fact, the study of Brodeur et al. (Unpacking p-hacking and publication bias, AER 2023) suggests that the peer-review process in Journal of Human Resources diminishes the p-hacking troubles that authors conduct.

**What is the basic logic in testing for publication bias (estimates v standard errors)?**

(1) linear models using the relationship between the estimate and its standard error

We always use the **funnel asymmetry test** discussed by Stanley (2005), because it has been shown to perform well in Monte Carlo simulations and is very intuitive. The test is based on the realization that in the absence of publication bias there should be no systematic relation between estimates and their standard errors. The authors of primary studies usually report t-statistics for their estimates, which means that they assume that the ratio of the estimates to their standard errors has a t-distribution, which in turn implies that estimates and standard errors should be statistically independent quantities. If, on the other hand, researchers prefer to publish estimates with a particular sign or statistical significance, estimates will be correlated with standard errors. The regression is heteroskedastic, so weighted least squares (with inverse of the variance as the weight) should be used. Btw, funnel asymmetry test is a way of testing the **funnel plot** of Egger asymmetry (you just take the inverse of precision and you switch the axes of the funnel plot and there is your regression of funnel asymmetry test): no relation btw estimates and standard errors means the plot is perfectly symmetrical. If publication bias is present, the funnel is asymmetrical (when the bias is related to the sign of the effect, so for example you are missing the part of the funnel with positive estimates of price elasticity because the law of demand says this elasticity should be negative) and hollow and wide (when the bias is related to the significance of the effect, so you could be missing some t-statistics that are). The effects from (esp. small) studies can be caused by publication bias but also by other factors. But this is just a graphical representation of the relationship: eyeballing a funnel plot is subjective, so funnel plot asymmetry tests are recommended instead. But funnel plot is a nice picture that serves for a good motivation, so please, keep it in your thesis. Note that the regression of reported point estimates on their standard errors is also called Egger regression (Matthias Egger, professor of clinical epidemiology at Bristol) thought it was actually Card & Krueger 1995 (David Card, Nobel prize in economics 2022, Berkeley) published in AER who first used the regression!!

If possible, in funnel asymmetry testing, researchers should use study-level fixed effects and cluster standard errors at the study level. Why study-level fixed-effects addresses partially the endogeneity issue of funnel asymmetry testing? It filters out the idiosyncratic study-level variation. The identification in the fixed-effect estimator, nevertheless, rests on larger studies (those that report more estimates) which could come as less intuitive. So if half of your studies have just one or two estimates and the other half are large studies with tens of estimates, fixed-effects would possibly not work well. Complement it with the study-level between-effect estimation, this accounts for the differences in the study size. You can also provide two weighting schemes of the funnel asymmetry test estimated by simple OLS: the one that weights by the inverse of the standard error which gives less weight to less precise estimates (and has a convenient advantage of eliminating heteroskedasticity from funnel asymmetry test) and the one that weights by the inverse of the number of observations reported in a study which gives the same weight to studies irrespective of how large they are (so that studies reporting large number of alphas do not get more weight). Always do the instrumental variable estimation where the instrument is the inverse of the square root of the number of observations the study uses (see the points referring to endogeneity in the funnel asymmetry testing).

These tests are all linear but the relationship really does not have to be linear. We simply do not know beforehand which model will be the best one. There are some intuitive reasons why linearity might not work. Say, the researcher gets an estimate that is just around the critical value of 1.96: a small increase in the standard error could deem the estimate completely unpublishable in the eyes of the researcher. There is nothing linear about the jumps at the conventional critical values of statistical significance. Thus, we have some tests that do not assume the linear relationship between the estimates and their standard errors.

(2) Non-linear tests using the relationship between the estimate and its standard error

**Stanley et al.** (2010) suggest that computing a simple average of the 10% of the most precise estimates only (and dropping the remaining 90% of the estimates with lower precision) corrects well for publication bias. This is often called Top10 method, it is straightforward but also overly simplistic. We often do not report it but if it serves its purpose, why not... **Ioannidis et al** (btw yes, we are talking about John Ioannidis, one of the most cited academics in the world, a famous Stanford epidemiologist) have used another method, also very simple one, they propose a correction procedure that focuses on estimates with statistical power above 80%., ie drop the estimates that have lower power than 80% and weight the rest by the inverse of variance. So you may see that both of these methods (Stanley et al and Ioannidis et al) only perform well if there is no heterogeneity in the data which happens, well, never. So, although you should get rid of most of the publication bias, you also get rid of a lot of variation in the estimates (in case of Top10 you are throwing away 90% of the info). The method of **Furukawa** extends the approach of Stanley et al. (2010) and suggests a way how to optimize the trade-off between getting rid of the bias and throwing the variability of data. He suggests to use only the stem of the funnel plot (not 10% of the most precise estimates but part of the peak of the funnel plot), and determines this portion/stem by minimizing the trade-off between variance (Stanley 2010: variance is lost when decreasing the number of estimates included) and bias (Stanley 2010: bias is increased when we increase the number of imprecise estimates included). It is quite clever and elegant, given that it is not a selection model. **Bom & Rachinger** account for the case that estimates get reported only when they cross a particular precision threshold. In their method they estimate this threshold and introduce an “endogenous kink” to extend the simple Stanleys (2005) funnel asymmetry test of publication bias. That is, the relation between the standard errors and publication bias is not linear nor quadratic in this test but the test tries to identify the precision threshold below which the relationship ceases to be linear in order to fit the data best. This means, that funnel asymmetry test is a special case of Bom & Rachinger => if the model identifies there is no kink, the relationship is only linear, and the kinked model becomes funnel asymmetry test. Selection is a linear function of the standard error for imprecise estimates and there is no selection for precise estimates.

Selection models assign weights to estimates to take into account that some studies or estimates are less likely to be published than others (note that the previous tests were not selection models). Selection models assume some distribution of the estimates and also look at the real distribution of the estimates and compare these two distributions. For example, statistically insignificant estimates will most likely receive a larger weight than significant estimates to compensate for insignificant estimates being less likely to be published. These weights are then taken into account when meta-analyzing the studies using a conventional meta-analysis model. The most well-known selection model is **Hedges** model and is frequently used in psychology and medicine. The method creates intervals based on p-values, and then estimates the weights for the studies with p-values belonging to these intervals. Studies in the same interval get the same weight. In economics, the most prominent one as of now (and it is implied you will report it) is the **Andrews & Kasy** selection model (=> btw yes, that Isaiah Andrews, the Harvard prodigy that got the prestigious John Bates Clark Medal in 2021 which basically sets path to Nobel in economics--and this is huge for meta-analysis) and the method works on the same principles as Hedges. It is not entirely clear what is the difference from Hedges and A&K even do not cite Hedges even though the methods are clearly very similar. A&K show how the conditional publication probability (the probability of publication as a function of a study’s results) can be nonparametrically identified and then describe how publication bias can be corrected if the conditional publication probability is known. The underlying intuition involves jumps in publication probability at the conventional p-value cut-offs. But there is a strong assumption involved in their selection model: the estimates and their standard errors are statistically independent. And of course, you don't observe the latent distribution of the estimates and their standard errors absent publication bias but only the resulting distribution distorted by publication bias.

So there is a test introduced by **Kranz & Putz** (2022); from our correspondence with Sebastian Kranz I understand that it was Isaiah himself who suggested the idea for the test---the idea is to compute the correlation between the logarithm of the absolute value of the estimated inverse elasticity and the logarithm of the corresponding standard error, weighted by the inverse publication probability estimated by the Andrews & Kasy model. If the model's assumptions hold, the correlation is zero. Note that this is a joint test of all the assumptions of the Andrews & Kasy model, not just the conditional independence of estimates and standard errors. For example, a strong assumption used by Andrews & Kasy and related selection models (but not meta-regression models based on the funnel plot) is that all estimates within a given group (e.g., negative estimates insignficant at the 5% level) share the same publication probability. Interestingly, we find for most of our previous subsamples a substantial correlation (not even close to zero), which suggests that some of the assumptions are violated. So for most (if not all) of our previous datasets the Andrews & Kasy model is not specified well.

On the other hand, it may be that the other techniques are also not specified well and their assumptions are just more difficult to test because they are laid out more vaguely than those of Andrews & Kasy (2019). But if you give an explicit rejection to the Andrews & Kasy model you just have to provide the reader with more alternatives in the analysis of publication bias. This is why we often use so many tests for publication ias, because they are quite different in their underlying assumptions and about how the publication bias could work.

While under publication bias, we understand the revealed preference of the authors or editors to publish or not to publish some specific type of results, p-hacking denotes the researchers effort to engineer results that are statistically significant by constantly trying different specifications. Although conceptually different, p-hacking stems from publication bias. **Elliot et al**. (2022) in Econometrica assert that the distribution of p-values in a meta-sample should have certain testable restrictions if there is no p-hacking in the sample. The paper relies on the conditional chi-squared tests by Cox & Shi (2022). The test for non-increasingness follows from the assumption that if there is no p-hacking in the sample, the distribution of p-values (usually below 0.15 but you can choose the value, do not go above 0.2) should be non-increasing. The test is histogram-based and thus you have to specify the number of bins such as 10, 15, or 20 (it should depend on how large your sample is but Elliot et al. do not say how you should choose the values and the trouble is that the results may differ substantially across the bins: I recommend to choose various numbers of bins and show the results are robust in all cases). Another histogram-based test of Elliot et al. is the test for monotonicity and bounds: the idea is that the p-curves based on t-tests are completely monotone in the absence of p-hacking, and their magnitude (and the magnitude of their derivatives) are restricted by upper bounds. So what you test again is the null hypothesis of no p-hacking.

The newest method comes from psychology, Frantisek Bartos and Max Mayer combine selection models and PET-PEESE using Bayesian model-averaging to adjust for publication bias (https://psyarxiv.com/kvsp7/). **RoBMA** tests for how well these different methods fit the data (A&K, FAT-PET, etc). Then some measure of weight is constructed, and the final estimate is the weighted average of these methods. But how the weights are constructed is, to be sincere, an unknown to me. But I do suggest trying RoBMA anyway.

**In the funnel asymmetry test and many others, you identify the publication bias assuming something about the distribution of the collected estimates. But those estimates vary immensely, why should they even fall under one distribution?**

In the funnel asymmetry test we do not have to make strong assumptions about the distribution of the parameter. We collect the estimates for which the authors report t-tests, and thus they assume that their coefficients and estimated standard errors are independent of each other (the ratio of these statistics has a t-distribution). This implies that the funnel plot will be symmetrical because there should be no relationship between the estimates and their standard errors.

This is, of course, conditional on the assumption that the standard error is exogenous to this estimate of the effect, which does not happen that often - that's why we try to control for data and methodology used in the study. Even so, it is likely that we are not able to filter out all the dependencies between the estimates. Especially, if we are not able to code for some methodological issue that would affect the coefficients and their standard errors in the same direction – our estimate of publication selectivity would be biased as well. The solution would be to use an instrument for standard error which would not be correlated with the methodological choice and such an instrument might be the number of observations used in the original study.

Anyway, we do not have to assume that the estimates are normally distributed. Our basic assumption that estimates should not be correlated with standard errors arises from the default assumptions, which are made by the authors of the original studies. Of course, it is possible to model the publication bias at the micro-level (to take the original datasets from the primary studies) but in economics we usually do not have the data at disposal. There are also structural models for publication bias but they usually do not work in practice, because we would have to strictly specify the mechanism by which this bias is driven.

The simple linear methods for publication bias cease to be statistically valid if the authors of the original works do not assume the symmetrical distribution of the estimates. This happens in some structural models but in econometrical works which estimate the parameters this is rather an exception and rather a t-distribution is present (or other symmetrical distribution which implies symmetrical funnel plot). Even for most Bayesian methods it can be shown that the posterior distribution should by asymptotically symmetric. Nevertheless, I fear that the assumption of symmetric distribution of a certain parameter practically guarantees a bias of the whole literature, because it is in a way a formalized publication bias.

**Is there any way to analyze the publication bias without assuming anything about the estimates and their standard errors?**

Most techniques assume that in the absence of publication bias there is no correlation between estimates and standard errors: meta-analysis has its origins in medicine, where the exogeneity of the standard error is rarely questioned (you can hardly have two medical trials with overlapping samples). In economics, however, the standard error can be endogenous for three reasons: it is itself an estimate (measurement error), publication bias may work through reporting artificially high precision (reverse causality), and some unobserved method choices may systematically influence both the point estimate and the corresponding standard error (omitted variables). No technique commonly used in economics meta-analyses allows us to get rid of the assumption. We can employ study fixed effects, which filter out between-study differences, likely the most important source of endogeneity. We can also employ the number of estimates as an instrument for the standard error, but some method choices can still be correlated with the size of the data set in primary studies (although it rarely is the case).

A more fundamental solution is provided by psychology, where the newly developed puniform\* technique (van Aert & van Assen, 2021) analyzes the distribution of p-values instead of estimates and standard errors. The foundation of p-uniform\* is the statistical principle that p-values are uniformly distributed at the mean underlying effect size: that is, when testing the hypothesis that the estimated coefficient equals the underlying effect. The idea of **p-uniform**\* is to find a coefficient at which the distribution of p-values is approximately uniform; this is done by recomputing the reported p-values for different possible values of the underlying effect and then comparing the resulting distribution to the uniform one. Following this principle, the technique’s test for publication bias evaluates whether p-values are uniformly distributed at the precision-weighted mean reported in the literature.

P-uniform\* comes from this **p-curve** concept, which is quite smart in its simplicity: take the estimates up to the 5% level of significance only, how does the distribution of p-values look like? If the curve describing the distribution of p-values is flat, there is possibly no effect in the literature. Right-skewed distribution (skewed towards smaller p-values) shows there is some effect, because there are more p-values concentrated up to eg 0.1 and 0.2, but much less so above 0.3. Left-skewed distribution (skewed towards 5% level) shows trouble: it means, that some p-hacking is going on. P-hacking means that you are adding observations one at a time to get to the statistical level of significance you like. Once you reach 5%, a level acceptable by most of the academia, bingo.. That’s how you concentrate p-values around the 5% threshold.

**Caliper tests** are typically employed to identify a systematic break related to publication bias at a particular psychologically important threshold (such as 0 for the point estimate or 1.96 for the t-statistic). The essence is to compare the number of estimates just below and just above a particular threshold: given a sufficiently narrow caliper, there should be no difference. Caliper test came to meta-analysis from political sciences; in economics, it appeared first in 2018.

In economics, instrumenting the standard error is possibly the easiest solution. We can employ the number of observations in primary studies as an **instrument** for the standard error. When the choice of method systematically affects both the magnitude of the estimated effect and its standard error, the explanatory variables in our meta-regression could be correlated with the error term. Why do we use the number of observations as an instrument for the standard error? Studies with more observations yield more precise estimates (so the standard error is correlated with the number of observations by definition), but the number of observations is little correlated with the choice of methodology.

**Weighting by precision is inappropriate in economics because some methods underestimate standard errors.**

Meta-analysts often use precision weights to remove heteroskedasticity in the regression estimating publication bias. If we find no evidence of publication bias, we can exclude the standard error from the equation and do not have to weight the estimates. But it is true that the assumption of exogeneity lies at the core of some most profoundly used tests of publication bias. We can partially address this problem by using study fixed effects, caliper tests, p-uniform\*, and by employing the number of observations in primary studies as an instrument for the standard error.

**How does the publication bias look like eg in the literature on the elasticity of intertemporal substitution? Getting negative elasticities is evidently wrong (it is the same as if you get a negative income elasticity of demand for a normal good or a negative price elasticity of demand for a Giffen good, these are wrong to obtain by definition).**

Suppose, for example, that a researcher estimates a negative elasticity of intertemporal substitution. A negative EIS implies convex utility, so the estimate is probably a statistical artifact. One should get negative estimates from time to time when the underlying EIS is small or estimation is imprecise, yet it makes little sense to build conclusions on them. There could be noise in data, there could be wrongly chosen estimation method, wrongly defined model, there could be small sample bias, many things can get you to the negative elasticities and even those do not necessarily have to be “incorrect” per se. The problem is that no upper limit exists which would mirror the lower limit of zero given by the theory: if many researchers discard negative estimates but most report large positive ones, our inference from the literature as a whole gets biased.

**Meta-analysis compares apples with oranges.**

Meta-analysis in economics examines heterogeneous estimates. Different estimates are produced using different methods, and you should try to control for the differences in the design of primary studies. If possible, provide separate results on subsamples. To increase the comparability of the estimates in your data set, focus on more narrow topics. Worst case, transform the effect to some common metric (such as the partial correlation coefficient which is not directly interpretable but keeps the ordinality of values so that one can say the effect is big enough or small enough to matter). The best common metric available is one standard deviation of an effect (which is easily computable if the original study provides summary statistics on their data).

**Studies of low quality should be excluded.**

Not necessarily. As Tom Stanley suggests it is always better to err on the side of inclusion. If you take into account only good studies, the obvious problem is where to draw the line between good and bad ones. We prefer to include as many papers as possible and give weight to different aspects of study design according to what we believe is the consensus on best practice. In this way we can explore the influence of different methods on the estimated effects. We also always try to control for the impact factor of the publication outlet and for the number of citations each study gets. This way, we do not throw away information from the unpublished working papers but we do control for this aspect in meta-analysis.

Some topics tend to be quite exhaustive, data collection takes years, there are many published papers. In these situations, it might be advisable not to collect the estimates from unrefereed materials just to make the collection feasible.

**The analysis omits some studies.**

We try to include as many studies as possible but may still miss some. We believe it is not a problem to miss some studies, as long as their results do not differ systematically from the results of the studies included. But we do publish a specific search query and datasets to allow for replication.

**Studies reporting many estimates dominate the meta-analysis.**

When each estimate gets the same weight, the unbalanced nature of data in meta-analysis means that studies with many estimates drive the results. One remedy involves the mixed-effects multilevel model, which gives each study approximately the same weight if the within-study correlation of the estimates is large. The problem is that the method introduces study-level random effects, which may be correlated with explanatory variables. With so many explanatory variables defined at the study level, we prefer to simply weight the regressions by the inverse of the number of estimates reported per study.

**Authors’ preferred estimates should get more weight.**

Studies examining the effect usually present many estimates, and often prefer a subset of these estimates (many results are shown as robustness checks). Some authors make it clear what their preference is, but for many studies it is impossible to select the preferred estimates. We control for data and methodology instead, which is easier to code and should capture most of the authors’ preferences, for example, the control variables in the primary studies. This way, we can also show how the methods that are not preferred but used as a robustness check might skew the results.

**Mistakes in data coding are inevitable.**

The collection of data for meta-analysis involves months (sometimes years) of reading papers and coding the data. We cannot exclude errors, but we do our best to minimize their number (either involving more people to double check parts of the dataset or having the collector to double check some random parts of it). You still have to keep in mind that this error rate in data is inevitable in any dataset of any empirical work and is present even in databases with such precise information as that from the firm’s financial statements. Unless there is no systematic bias these errors could potentially create, we should be fine.

**Individual estimates are not independent, because authors use similar data.**

Meta-analysis was originally designed for synthesizing medical research, where individual clinical trials can be considered approximately independent. It is, to some extent, similar in experimental economic research. In contrast (in observational research in economics), most regression results reported are not independent, but neither are the observations in most economics data sets. To account for the dependence among observations we cluster the standard errors at the level of individual studies, and/or data sets (say from the same databases)…

**Why model averaging methods are used in heterogeneity analysis and how does it work?**

In meta-analysis, one uses many explanatory variables describing the design of studies. But how can you be sure that controlling for the cross-sectionality of data or for the number of citations is really important in this literature? Quite probably, many of these variables we code are redundant. This is called model uncertainty: we are not sure, how the best model should look like because there is just no solid theory to defend our choices.

The general-to-specific method (throw away the insignificant variables one by one) can be always used but we do not consider it to be statistically valid because t-tests are not designed to be run conditionally on each other. Ridge and Lasso do similar thing. So we do not want to throw away the variables willy-nilly but we do not want the potential redundant variables to skew our results (this could inflate the estimated variance of our parameters). BMA (Bayesian model averaging) estimates models designed with many different subsets of all the explanatory variables on the right hand-side. So we have tens of thousands of models and what this technique does is that it weights each model by, let’s say, a goodness-of-fit adjusted to model size (something like adjusted R-squared in frequentist econometrics), and produces the weighted mean coefficient for each of the variables including the posterior inclusion probability, which says how likely is the variable to be included in the model. The Bayesian part of the model averaging works with priors on 1) estimated coefficients and 2) models.

**What are these BMA priors?**

You have some ex-ante preference on how the model or the coefficients should look like or behave. There are two key selections for priors:

**The first one is the g-prior (selection on coefficients)**, which roughly indicates how much weight should we put to our prior or preference for all regression coefficients to be zero. Usually we use **unit information prior**, which gives to this prior the same weight as one observation in our data (meaning the larger data set you have, the smaller weight the prior has). Then there are **plenty of modifications** - some believe that these modifications have slightly better forecasting power (like BRIC prior, hyper-g prior, etc.), but basically the choice of prior seems to be of a small importance, because it does not have any influential impact on our BMA results (at least according to what we have tried in our meta-analyses).

**The second one is the model prior (selection on models)**, which roughly indicates how much prior weight a certain model has (like in BMA it would be a model with some subset of all potential regressors). Usually, people use the **uniform prior**, which gives equal weight to each model (because in meta-analysis, there is usually no reason to prefer some model over other model). But if you have certain number of regressors, say K regressors, the largest number of potential models has K/2 variables. So it would seem that if you want to use this kind of prior, you implicitly want the best model to have K/2 variables, which is usually just too many. The alternative might be this **random prior** (which by the way often has different results but might be a better choice), which gives the same weight not to each model but to each model size (same weight to each number of variables that we include in a model). On top of the uniform model prior we usually use the **dilution prior**. In this prior the relative weight of each model is further multiplied by the determinant of the correlation matrix of the variables included in the model. The dilution prior is designed to address collinearity: models with high collinearity will have small determinants of the correlation matrix, and therefore little weight in our implementation of BMA. This is quite important for meta-analysis because datasets often reek of multi-collinearity (which we cannot ignore but often put less weight on if the model is highly stable – meaning it is very robust even when these problematic variables that create multi-collinearity are thrown away).

**The analysis omits some factors that may cause heterogeneity in the reported estimates.**

We usually collect tens of aspects of data, methodology, and studies that may affect the estimated effects. It is often so that more specifics of study design could be included. But you have to draw a line somewhere for the data collection to be feasible. Still, we often collect more variables than most meta-analyses in economics. Nelson & Kennedy (2009) review 140 meta-analyses and report that a median analysis uses 12 explanatory variables; the largest meta-analysis has 41 variables. In 2021, the largest meta-analysis in the sense of using many explanatory variables was Gechert et al. (Review of Economic Dynamics) using over 70 variables. But note that for this number of explanatory variables you have to have a huge number of observations at disposal.

**What about weighting by standard error in a meta-regression?**

It is a matter of taste whether to use weighted least squares in meta-analysis (to weight all of the explanatory variables including the effect in question by precision) when other explanatory variables than the standard error are included. Tom Stanley argues to always use weighted least squares, because of the heteroskedasticity problem and because weighting always gives priority to more precise results (you have to realize that Tom Stanley had the original idea of using WLS). I prefer not to weight the regression by precision if the regression contains variables defined on the study level, like the number of citations. Because precision differs for each estimate within a study, weighting by precision introduces artificial variation in these variables. Since both approaches often yield very different results, it might be a good idea to report the results of the other approach as a robustness check only, helping (especially) to identify whether the publication bias survives inclusion of other aspects of study design in your meta-regression.

The story is different when you weight your data by the number of estimates per study. Now you are giving each study the same weight, meaning the large studies do not drive your results.

I do not think there is something inherently good or bad about any of the approaches. The “clean and neat” way, which is my opinion, is to show what the unweighted regression yields, since you are already controlling for the idiosyncrasy with the coded explanatory variables. Of course, you cannot bury the reader with hundreds of different results and robustness checks, choose one you find is the best and leave the others for your memoirs (appendices can take anything).

**Meta-analysis may disagree with large primary studies.**

The major reason for conducting meta-analyses in medical science is to increase statistical power by combining small but costly clinical trials. Because individual clinical trials use similar methods, a comparison of a meta-analysis with a later, large clinical trial provides a viable test of the reliability of the meta-analysis. In economics the methods differ, and meta-analysis can be thought of as a weighted average of many different approaches. It would be difficult to construct a primary study reflecting all recent advances in the methodology related to our topic/meta and all possible aspects of our definition of best practice. Moreover, the advantage of meta-analysis is the ability to evaluate the systematic effects of various method choices, and discuss the consequences of changing the definition of the best practice. Unlike primary studies, meta-analysis does not rely on a particular data set. We can evaluate the mean effects of the choice of different methods (the effects may differ for different data sets, which is unobservable for the authors of primary studies. For example, a combination of panel data for a distance of trade measured in minutes is a dataset impossible to get and to be used in a primary study but meta-analysis can control for such combination of data choice and infer conclusions based on such choice), and thus hope to obtain more robust results.

**Couldn’t we just replicate the old primary studies?**

Yes and no, both replications and meta-analysis have their merits. You probably heard already a lot about the **replication crisis** not just in the economics but more general in sciences. And many well published papers talk about the evidence that what we see published in our journals may not always be representative of the underlying data and findings. That despite following the same protocols, replications of published experiments frequently find different effects than those from the primary studies. There are two important explanations to this phenomenon:

* One of them is **heterogeneity**. This is context dependence. Say you measure a utility function parameter based on the sample of people from one village, but doing the same experiment with people from the other side of the planet, you may get completely different result. So there is a lot of context dependence in economics and no universal rules valid for the entire population.
* And the leading explanation for replication failure is **publication bias**. Findings that have certain characteristics which make them get published more easily, like statistically significant findings, findings more consistent with a theory, even findings that are surprising or considered provocative, and maybe more likely to be published than null results or seemingly boring results. Those, that make for a good story.

And publication bias has become a topic of many research articles especially more recently (say since 2018) by researchers from Harvard, MIT, Stanford, there is a paper in American Economic Review (Isaiah received his John Bates Clark medal by the American Economic Association for his work on /among other/ publication bias, and is a holder of the MacArthur Genius Award) and Econometrica (Elliot et al) dedicated to a new method of treatment for publication selection bias and a paper from MIT (Furukawa) which stresses that publication bias does not necessarily have to be about people wanting to give biased information but about people trying to be efficient in presenting their results, thus not reporting all the results, thus committing to a bias.

The fact is that public policymakers, foundations, citizens, they look at published research literature for guidance on many issues, especially those policy related. But if they see really a selective subset of the research that was done, the question remains - how much information do we really get from the published empirics. Meta-analysis is one of the responses.