**Can I have more than one search query?**

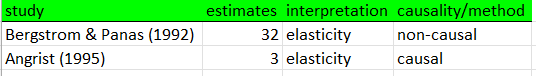
The search query is provided so that other people can replicate your search of primary studies (the original studies that report the effect in question). Of course, if you use snowballing (checking bibliography references from the recent papers that do estimate your effect), your data search becomes hard to replicate. If you can help it then yes, having only one search query would be ideal. If one search query is not sufficient for the literature you deal with, just provide an appendix with PRISMA diagram and searches you used. Use only Google Scholar, no need to complicate your searches elsewhere. From your query (using the recommendation of checking the first 500 items), you might download around 200 studies. From those 200 studies only a portion would be eligible for use in your meta-study.

**How many studies should my meta-analysis include?**

Every literature is different. If your literature deals with Mincer wage equation, for example, one would expect many studies with many observations. How many is many? Median bachelor or master thesis on meta-analysis uses around 60 papers that involve some 1,000 observations (estimates of the effect). Small datasets have around 40 papers with 600 observations. Such meta-analysis has less degrees of freedom but still can make sense. Collecting effects from only 20 studies, for example, you have almost no between-study heterogeneity so the literature is just not ready for a meta-analysis yet (note that in medicine, this would be relatively large met-analysis; in psychology and economics, however, the practice is different). The larger dataset, the more impressive your meta-study is. Bottom-line, you should collect ALL the relevant estimates available. If you think the collection is not feasible (there could be hundreds of studies and thousands of estimates), let’s discuss what subgroup of studies could you focus on. Include all studies from all the previous meta-analyses as well.

**Interpretations of the effects I want to collect are vastly different, how should I decide on which effect to collect?**

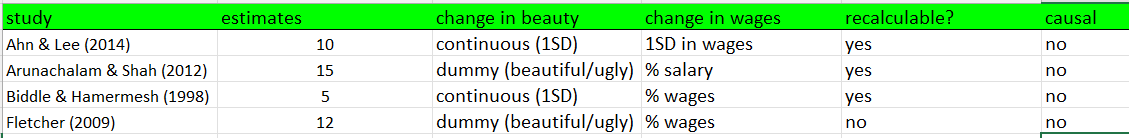
Say you have some 200 studies that estimate similar effects (effect of beauty on wages). Open your excel spreadsheet, go through every single study and make a note, eg:



If elasticity is the effect in question, your job is easy, especially if the elasticity is the coefficient that is directly estimated (elasticity is not estimated as an inverse of a regression coefficient, for example, that would completely skew your publication bias tests). It gets harder with estimating marginal effects where the left-hand side of the regression as well as the right-hand side of the regression have different variables involved. Always make sure you understand the interpretation of the effect the author uses.

* + The explanatory variable can be continuous (such as height of a man). It could be the interpretation is the effect of 1 cm of additional height is associated with some %-change in wages. It can be the effect of 1 inch of additional height associated with one-standard-deviation increase in wage.
  + The explanatory variable can be a dummy variable (such as “high person” = 1 if height of a man is more than 180 cm). Then, the interpretation changes: people considered high earn some % more in wages than those that are not considered high.
  + Also, the left-hand side of the equation, say wages, can be in absolute terms (say in dollars), or can be in logarithm (if the researcher wants the interpretation to be relative, such as % change in wages), or can be an hourly wage instead of yearly household income.

Thus, it matters utterly you compare the effects that are comparable, and you can recalculate it to a common metrics. If this is the literature you face then again, create a spreadsheet and a quick summary of what is available in those 200 studies you have at disposal, such as



When you know roughly how many estimates from how many studies you get that are well comparable then you can decide how to proceed further with your data collection.

**Other helpful tips on data collection**

* Do not code immediately OLS as a dummy, VAR as a dummy, and so on. Create one column called “method” and make a note on which method is used. Same goes for coding for geography or other group variables that are exclusive to each other. You will decide later which variables you will use for your analysis (maybe only methods controlling for endogeneity and those that do not control for endogeneity, and maybe only developing and developed countries). The columns could look like this:

Table

Description automatically generated with medium confidence

* while collecting data, you will be reading a lot. If you see something interesting, a relevant argument, a relevant discussion, copy-paste it to another file, let’s call it “notes”. Redistribute these notes regularly into different categories, say notes on “how studies deal with endogeneity”, notes on “ideal experiment”, notes on “how to define short and long term effect”, notes on why it is important to “differentiate between male and female subjects”, notes on why surveys are full of “measurement errors”, notes on specific “biases in the literature” such as attenuation or omitted variables bias.., why “time-series” data yield better estimates than “cross-sections” in your literature…

**My dataset does not really have many marginal effects, I mostly deal with odds ratios and risk ratios. How to deal with such dataset?**

* check the wonderful thesis of Daniel Bartusek on how to handle such data at

https://dspace.cuni.cz/handle/20.500.11956/126499

**I have trouble finding a reasonable number of estimates that can be translated into common metrics.**

Most trouble we face is because the authors in primary studies use continuous as well as discrete variables to estimate their effect. The interpretation of effects based on such measures is different. To be able to compare the effects, nevertheless, you need some common metrics. There are two ways to deal with the problem:

**1) one is to use the effect of one standard deviation** of a variable people use. For example, the authors might use a continuous variable identifying how beautiful people are (in the literature dealing with the effect of beauty on earnings, for example), such variable comes from surveys that ask people how beautiful on a scale of 1 to 5 they think a person is, so the variable is continuous, has a mean and a standard deviation. If you know the standard deviation of the sample from a primary study, you can simply estimate the effect of one standard deviation change in beauty scale on earnings.

If the authors do not use a continuous variable but a dummy variable, for example very attractive = 1 if people are rated better than 4 and ugly = 1 if people are rated below 2, the variable is discrete. But if you have the original distribution of the attractiveness measure you can still estimate the standard deviation of the distribution with ease. True, the authors rarely report distribution of the original continuous variable while using discrete variable instead.

So one way out of it is to recalculate all that can be recalculated to a common metric, such as the effect of one standard deviation change in a variable. Or you will consider only those subsamples that have the same interpretation, sacrificing the loss of some of the estimates of the effect. Another option, a nuclear one, is

**2) to use partial correlation coefficients**. This metric is used when there really is no reasonable way to put together differences in definition of the estimates (say academic achievement can be measured as credits earned or as a probability of graduating from high school or as paying attention during classes and you have not really much information on the distribution of these variables in the primary studies). The translation of the estimates is difficult or questionable or impossible. Then you can use partial correlation coefficient. But know that this is your last option and **try to avoid it like a plague**. The PCC and the corresponding standard error formula is following:

PCC = t-statistics / sqrt(t-statistics squared + degrees of freedom),

SE(PCC) = sqrt{(1-PCC^2)/degrees of freedom}.

**Sometimes authors in primary studies report only t-statistics and not standard error. Sometimes, they report standard deviation instead of standard error. Sometimes they report 95% confidence intervals. Sometimes they report p-values only. Sometimes they report only signs for the level of significance, like \* for 10% or \*\*\* for 1% level. Can I use those statistics to get to the standard error? What if there is no measure of precision reported?**

1) Only t-statistics is reported. One needs to be aware of what hypothesis the authors are testing when they report these standard errors and t-statistics. The null hypothesis is that the regression coefficient equals zero (B0 = 0). That is,

t-statistics (Best) = (Best – B0)/SE(Best) = Best /SE(Best).

So yes, you can calculate the standard error by simply dividing the estimated coefficient Best by its t-statistics.

2) Only standard deviation is reported. You can easily calculate standard error using the square root of number of observations (you should collect the number of observations anyway):

SE(Best) = SD(Best)/sqrt(number of observations)

3) Only confidence intervals are reported. Say you have a mean effect of B = 0.05 with 5%CI = -0.12 and 95%CI = 0.23. Then, the standard error of B can be calculated as:

SE(Best) = (95%CI - 5%CI)/(2\*1.96)

In our example, [0.23 – (–0.12)]/(2\*1.96) = 0.089.

4) Only p-value is reported. Use it to calculate t-statistics, for example via MS Excel function TINV:

t-statistics (Best) = TINV( p-value \* (3 - #sides), nobs),

where p-value is something like 0.01569, # of sides should be always 2 if the paper makes sense in reporting their 2-sided p-value, nobs is the total number of observations = cross-section observations \* time series observations (say 352), thus t-statistics=TINV(0.01569\*1,352)=2.43.

5) Only star-signs indicating significance levels are reported. Possibly, the only measure of significance you get reported is \* for 10% level of significance, \*\* for 5% and \*\*\* for 1% of significance. This is very tricky because unless you contact the original authors and ask them for the numerical results you will introduce some error to your data one way or the other. You can approximate the 10% level relatively easily by choosing a value of t-statistics between 1.645 and 1.96. Here, the error you introduce is somewhat acceptable if you do not create this error consistently (if this is a case of few single instances). You can approximate the 5% level of significance by choosing the value for t-statistics between 1.96 and 2.575. Again, you cannot have too many of these approximations or you will introduce a systematic error to your dataset. The worst choice is to choose a t-statistics to approximate 1% level of significance => this can be anything south of 2.575 and I would choose rather not to include these estimates into your dataset or try using other additional information you have (like the number of observations) to approximate better the t-statistics. In any case, you should mention this in a footnote or an appendix dedicated to data adjustments and recalculations.

6) Reported value of p-value (standard error) is zero. First, contact the authors and ask them for their numerical results. If they do not respond, you have few options. If the value is say 0.000, you can decide to: (1) either calculate the t-statistics according to the formula above for an approximate p-value of say 0.0002 (so something between 0.0000 and 0.0004) noting that it is an approximation (the sample size can also determine whether it is a good or a bad approximation) or (2) you will ignore these values. Personally, if there are not “too many” of these zero p-values, I would rather go for erring on the side of inclusion. If you decide in the end to collect and transform these estimates, please do so ex-post, after the whole dataset is collected. The reason is: you may find out that these estimates are outliers in your dataset and will fade away with winsorization---thus there is no reason for any transformation.

7) No measure of precision is reported. If you wonder whether it even makes sense to collect estimates that do not report any measure of precision than the answer is yes, it does make sense. If you have no measure of precision of the effect, such an effect can still be used for the motivation of the paper, for example for the calculations of simple averages of the effect from different subsamples. Of course, in publication bias analysis you need the standard error, so you cannot use any such observation there. If you find, however, that publication bias is not present in the literature, standard error becomes obsolete, and you do not have to have it in your heterogeneity analysis (if you do model averaging or other technique capable of dealing with model uncertainty, nevertheless, try to always include it). Thus, the effect without standard error can be further used in the analysis of heterogeneity if publication bias is not present in the literature and you do not have to discard such observation. For consistency, I would prefer to have a full panel dataset so that all of my estimates have their standard errors. But if you find yourself with a small sample size then go for the estimates without standard errors as well.

The estimate I collect from primary studies is not the effect in question. The effect in question is a function of the estimate I collect from primary studies. How can I estimate the standard error of such an effect?

There are three kinds of transformations you may face, one that is made

1. using some constant (meaning that no new imprecision is coming from the transformation),
2. using a different functional form of your estimate, and
3. using some other estimate to transform the collected one (which involves another standard error entering the transformation).

In the first instance (1), you may need to make some trivial transformation of the effect in question. Say the effect of 0.03 means that an additional 1 inch of height is associated with 3% increase in wages. The effect has a standard error of 0.01 (thus t-statistics of the effect is 0.03/0.01=3). To transform this effect from inches to centimeters, you would multiply the effect by 1/2.54 (because 1cm =1/2.54 inch). But what about the standard error? Well, t-statistics of your original estimate is 3 and that´s a given, inputted constant does not change the precision of the original effect. So you have to make the same transformation for the standard error as well in order to keep t-statistics of a new estimate at 3. The new effect is 0.03/2.54 = 0.013, additional increase in one’s height by 1 cm is associated with 1.3% increase in wages. Standard error of the new estimate would be 0.01/2.54 = 0.0039. New t-statistics is the same as the old statistics, because transformation with a constant does not bring any new variation to the estimate.

In the second instance (2) and third instance (3), you may need to make some atypical transformation of the effect (see common examples below – that is involvement of interaction term or involvement of a squared term in a regression). Or: say you collect the estimate B1 but in reality, the elasticity B is estimated using 1/B1, or say you collect estimates B1 and B2 (each estimate has its own standard error, of course) and you need to estimate B = B1/B2. How can you estimate the standard error of such an estimate? We use something called “delta method” (if you have never heard about it, write to me and I will sent you some study materials). There is a formula in R which can make your life easier:

*deltamethod(g, mean, cov, ses=TRUE)*

where g is the transformation of the parameter for which we want to estimate the standard error, mean value of the estimated parameter and cov denotes variance-covariance matrix. In case of only one parameter, you just input a variance and the syntax is really easy. Say the authors estimate X, the elasticity is equal to 1/X, and say X=0.5 with a standard error of 0.2, then:

*install.packages("msm")*

*library(msm)*

*deltamethod(~ 1/x1, 0.5, 0.2^2)*

The function can be used even when the elasticity is a function of more parameters, just set the variance-covariance matrix in such a way that the covariances equal zero. Syntax is then:

*cov<-(diag(c(se(x1)^2, se(x2)^2), 2, 2))*

*deltamethod(~ g(x1,x2), c(x1,x2), cov)*

In any case, your publication bias analysis should be made on the original estimates collected from the studies or transformed values if the transformation does not skew in any way the relationship between the coefficient and its standard error. (An example of this would be: say B is the regression estimate with the standard error SE(B). The reported elasticities from primary studies are 1/B. This is a convex transformation, assuming B is bounded away from 0. The expectation of 1/B is therefore increasing in SE(B), holding constant the expectation of B. If you do the standard publication bias tests on the elasticity and find the publication bias present, it is exactly what you would expect to find. But the pattern is introduced by the transformation of elasticity not by selective reporting. In this case, you have to do your whole analysis on the original coefficient B not on the elasticity values = 1/B. Only after you do your analysis, to give your audience the economic interpretation of the effect you would make the transformation.)

**Sometimes the authors in primary studies report the effect and some interaction with the effect (for example, the effect in question is education, the regression reads “wage = A\* education + B\* education \* female”). Do I have to take the interaction term into account?**

The interaction term “education\*female” is directly taking part of the effect of “education” out of the picture, ignoring it would be a mistake. Let´s present an example:

wage = education + education \* female + …

= 0.2 + 0.4 \* 1 or 0 +… *=> coefficients*

(0.1) (0.3) (0.0) *=> standard errors*

If female =0, the effect of education for males is 0.2 with the standard error of 0.1.

If female =1, we use the delta method to estimate the effect and its standard error. The effect of education for females is 0.2+0.4 = 0.6, the standard error is a square root of the sum of the squared standard errors of both term and interaction term, i.e. sqrt(0.1^2 + 0.3^2). Note that this calculation is missing one very important information: the covariance between education and female terms. If you have information about the covariance (say the covariance equals 0.005), the correct way of estimating the standard error would be sqrt(0.1^2 + 0.3^2 – 2\*0.1\*0.3\*0.005). The authors of primary studies rarely (basically never) report the covariance term and you can thus approximate it inputting zero or something very tiny instead. Indeed, you are now aware that this calculation is just an approximation of the standard error (but mostly a very precise approximation).

**Sometimes the authors in primary studies report the effect and the effect squared (for example, the effect in question is education, the regression reads “wage = A\* education + B\* education \* education”). Do I have to take the squared term into account?**

Absolutely yes, the squared effect presents a quadratic element to your effect and is part of your effect. Let´s present an example:

wage = education + education\* education + …

= 0.2 + 0.4 +… *=> coefficients*

(0.1) (0.3) *=> standard errors*

Now, we need the sample mean of education variable, lets say 0.17 (ln years).

Then, if B1 is the effect of education term and B2 is the coefficient of education-squared, then the effect B of education is (using delta method):

* B = B1 + 2 \* B2 \* education(sample\_mean) = 0.2 + 2 \* 0.4 \* 0.17
* SE(B) = sqrt{[SE(B1)]^2 + 4 \* [SE(B2)]^2 \* [education(sample\_mean)]^2}

Covariance is again set to zero.

**Sometimes authors in primary studies report negative coefficient and positive t-statistics so that the standard error is negative. Is that even possible?**

This is not possible by definition (remember that SE is a square root of something). If the coefficient is negative, the t-statistic cannot be positive, unless the authors of the original studies report the t-statistic in absolute value. But we do not code for anything in absolute values, we have to take into account the positive or negative signs if our meta-analysis should make any sense.

Of course, it could also happen that there is a typo in the paper and the minus is just missing and needs to be added. It may be that the authors write that they are reporting standard errors and are in fact reporting t-statistics. If you're not sure where the truth lies, check with the authors themselves. In any case, it must be true that by definition a negative coefficient is associated with a negative t-statistic and a positive coefficient is associated with a positive one. The standard error is a positive non-zero number always and forever and ever. Anything else is a mistake and it is up to you to verify and correct it.

**Some examples of potential variables that could appear in a meta-analysis**

If there is any meta-analysis related to your topic, find out which explanatory variables they use. I provide here just a list of potential variables for inspiration.

variable definition:

* right-hand side variables, for example how education is defined: eg. education = years of schooling, education = attained academic degree, education = some grading system (GPA etc)..
* left-hand side variable, for example, how wages are defined: eg. wages = income per time, maybe it is rather a monthly consumption or yearly income (or wealth per person in a household), possibly productivity (test scores, inside-company rankings)
* could there be a short- and a long-run estimate? Note that this can be perfectly correlated with some of the methodological variables.
* Is this the author’s favored estimate (preferred estimate or the one that represents the main results of the paper or is this just another robustness check)?

data characteristics:

* number of cross-sectional observations (eg. how many people entered the experiment)
* number of time observations (eg how many years are in the sample)
* data dimension: cross-section, longitudinal (panel, time-series)
* sample mean year of schooling, sample mean year of experience (from the primary studies)
* source of data (do they use some specific database), data character (input is micro data, input is survey data from households or employers, input is from national register), could be the self-reported data are used in some case which are more prone to measurement error, how are the data aggregated (micro data, city data, regional data, industry data, country data, and so on)
* estimate aggregation (is the estimate relevant for a region or a city or a country so this could be coded as: region estimate, city estimate, industry estimate, country estimate, intergalactic estimate. Note that a regional estimate can be estimated on micro data.
* data year (log of the average year of data), first and last year used in the sample has to be collected
* data length (how many years is a study researching, time period in question)
* scientific field of primary study---economics, psychology, neuroscience...

experimental setting (relevant only for experimental literature)

* real reward, hypothetical reward (if the reward was really paid in money or not)
* framing (positive when one gets rewarded, negative when one gets punished-suffers a loss)
* real stakes (how large the reward is to median household expenditures thus saying how is the reward important for a person)
* measure of performance is qualitative, quantitative, or both
* nature of the experiment---judgement & decision, game & market,
* task: cognitive vs. manual work or appealing vs. non-appealing..
* type of motivation (altruism, trust, reciprocity, fairness)

spatial/structural variation

* type of education: kindergarten, primary, secondary, tertiary (so the data the researcher uses relate to specific grade)
* type of school: private, public, charter, placement, virtual... (possibly also some school quality measure)
* field of education: STEM, social sciences, medicine, humanities
* sector of employment: private, public, specific industry (manufacturing) or type (IT)
* employee: self-employed or dependent on employer
* gender: male, female, mixed sample (general public)
* family: dissolved, cohabitating, married, divorced, single, same-sex parents
* ethnicity: Caucasian, minorities (blacks, natives – Indian, Innuit), immigrants..
* age: adolescents or adults, sample mean age of respondents, younger/older group
* religion: Christianity, Islam, Hinduism, and Buddhism..
* geographical variation: which country or region, possibly developing/developed, urban or rural, mean income per capita or consumption per capita by country of employment (which should correlate with development so pick one)
* minimum wage = 1 if there is a minimum wage distorting the labor market (in years related to data sample)
* was there some schooling reform made during the sample period (this would possibly concern experimental studies)
* other group variables: was the effect estimated for some specific group of people? For example, only for working mothers, for those working while studying, part-time students, educated in a foreign country, low income such as those who have subsidized lunch, coming from dissolved families, students with learning disability, low-performing students, students with bad teachers (teachers from lowest percentiles in education or experience), dropouts from high-school or university vs. those who finished education, teaching quality (how good the teacher is), and so on. Note that this is a spatial variation so these are not the control variables from the regressions in primary studies. Working mothers group = 1 means that all workers in a sample are females with kids (the specificity comes either from specific dataset or from the inclusion of the interaction term). Working mothers control = 1 is specific of the estimation method (see below) and it means that the original regression controls for females with kids.

estimation method

* Mincer equation or full-discounting method (or model if you will, is it log-log, level-level, log-level equation..)
* OLS, difference-in-difference, matching method, etc (ad IV: could be interesting to check what are they treating by IV, if it is the measurement error they treat for or smth else, if they are instrumenting ability, if they are treating for endogeneity...)
* Experiment (note that experimental study can be estimated using OLS) - field experiment, lab experiment, lab simulation,
* or quasi-experiment (like IV, regression discontinuity, difference-in-differences, matching methods, and sometimes even panel methods which use unit-fixed effects with such design that the estimates can be interpreted as causal and not only correlational)
* Unit-fixed effects, Time-fixed effects
* ability control: direct control by IQ or a proxy, cognitive vs non-cognitive control, instrumenting ability by parental education or parental income (or other)
* other controls: age and age-squared control, experience, ethnicity control, health control, gender control, working mothers control, technological change (simple time control, this is not time-fixed effect), occupation control. firm-level characteristics, macro variables included..
* Are social returns also estimated or only private returns?
* controls for class size, teachers, the technology/material quality used in class
* funky controls: signaling effect, sheepskin effect
* On a general note: make notes on how authors treat endogeneity, measurement error, omitted variable bias (IV, natural experiment..) How do they assess causation (vs simple correlation). Do they comment other types of biases, like attenuation (least squares estimates biased in magnitude towards zero, see iron law of econometrics by Jerry Hausmann, attenuation usually stems from measurement error), how do they treat for ability bias, small sample bias, maybe publication or other selection biases...

Publication characteristics (please collect this information as the last one so that it is up to date, and if possible, collect it all in one day)

* Impact factor (<https://ideas.repec.org/top/top.journals.rdiscount.html>, <https://ideas.repec.org/top/top.wpseries.rdiscount.html>)
* Citations = log of number of citations per year in Google Scholar (ideally since working paper version first appeared but not necessarily).
* Published study = 1 if peer-reviewed journal publication, =0 if working/discussion paper or book or other
* Publication year = the year in which the study was published minus the minimum publication year of the sample

**I have collected the dataset, what to do next?**

1. check if there is any nonsense in the data, i.e. negative standard errors (which can be caused by a typo or copying a typo, either in standard errors or t-statistics---if something like this occurs, check the original paper, or write to the authors for an explanation), zero standard errors (it is not possible for a statistical estimate to have perfect precision because than it is a constant not an estimate - this may be due to the fact that the standard error was calculated from a zero effect, so it is a rounding error).
2. make a funnel plot, on the horizontal axis the effect, on the vertical axis the inversion of the standard error, have a look. Are there any weird observations that lie “outside” of the funnel? Check if any mistake was made while collecting these observations. If in doubt, send me a picture, we'll take a look at it together.
3. create a table with summary statistics: mean, median, weighted mean where the weight is the number of observations per study, min, max, take a look at the outliers, decide on winsorization (outlier is an observation that stands out like a huge maximum, for example). Winsorization is a way of dealing with outliers, some researchers rather throw away the observations that are suspicious, I suggest winsorization. Try winsorizing at 1% level first, see how much the mean changes from the median, then winsorize at, 2%, 2.5%, see if the mean changes "a lot". We want the smallest winsorization possible, which does not use too many observations but at the same time removes meaningless outliers in the data. We do not want to throw away the observations.

Any estimation you do, cluster the standard errors at the study level at least (you can also use clustering at the author level if you see many papers have similar teams of authors). With fewer than 30 studies I recommend wild bootstrap.