Forthcoming in: Stimson, R.J. (ed.) *Handbook for Spatially Integrated Social Science Research Methods*. Cheltenham UK: Edward Elgar, 2013.

Meta-analysis of previous empirical research findings

Jacques Poot

1. Introduction

As in other fields in the social sciences, the number of applied spatial studies on any given topic has been growing very rapidly in recent decades. This trend is not just the result of an increase in the number of academics and others actively conducting empirical research, but also because of path breaking changes in computer power and storage, the development of new methodologies and a "flood" of numbers on all aspects of life. All this research activity has become increasingly accessible through the internet with electronic publication of working papers, journal articles and, more recently, books as well. Search engines such as *Google Scholar* give the student and researcher instantaneously a list of recent studies on any topic. The scientific impact of each contribution can be readily, albeit imperfectly, gauged by means of the number of "hits" of a webpage, downloads of an article, or the number of times a paper has been cited to date.

Chapter 9 of this book discussed how a student or researcher can efficiently and effectively extract information from what is often a vast amount of literature on a topic in order to write the literature review. The literature review aims to be an objective assessment of what is known on a particular topic and, more importantly, may suggest what is not known yet. This can be the basis for formulating a new project: either developing new theory, or conducting new empirical analysis, or both. The literature review is commonly a narrative and qualitative assessment of the research that has been conducted up to that point, to the extent that the findings are publicly available. However, if the research is empirical, i.e. based on real world data, and concerns the statistical testing of a particular hypothesis or estimation of a particular parameter, it is in many cases possible to formulate a statistical model to explain differences in results between different studies. This kind of quantitative synthesis of previous findings in empirical research is referred to as "meta-analysis". Meta-analysis has been hugely increasing in popularity in recent years, particularly in experimental sciences such as medicine and psychology, but also in the non-experimental social sciences. A handbook chapter on this topic is therefore warranted. That is the purpose of the present chapter.

The term meta-analysis was first introduced by Gene Glass in his 1976 presidential address to the American Educational Research Association. Glass (1976) defined meta-analysis as follows: "Meta-analysis refers to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempt to make sense of the rapidly expanding research literature." Glass categorised research into primary analysis, secondary analysis and meta-

analysis. Primary analysis refers to an original study using a unique dataset. Secondary analysis re-uses the same data as the primary study, but does further work with it. This has several benefits. Firstly, it is useful to simply replicate the calculations of the original researchers to check one's software and programming. This may even reveal mistakes in the original research (e.g., Hamermesh, 2007). More commonly, researchers use secondary analysis as sensitivity analysis to see how robust the original findings are to changes in for example: sub-samples of data used; the variables considered; or the estimation methodology employed.

In the social sciences it is generally accepted that results are often context-specific, which implies that it is very hard to draw robust general conclusions. This has been particularly the criticism of empirical economics (for example, Leamer, 1983), but in recent years applied econometrics has again gained respectability by new methodologies that come closer to the controlled experiments of the exact sciences (Angrist and Pischke, 2010).

The word "meta" comes from the Greek language and means basically "beyond" or "about". Meta-analysis should not be confused with "meta-data" which refers to information about data. For example, the meta-data of a survey can often be found in a document that reports the number of people interviewed (the sample size), the way in which people were recruited (the sampling strategy), the questions asked (the questionnaire), how the information was coded (the list of variables), etc. We can say that meta-data refers to "data about data" while meta-analysis refers to "analysis of past analyses".

Meta-analysis has a long history that goes back to Karl Pearson's (1904) analysis of various estimates of the impact of vaccination against typhoid on incidence and mortality. The general idea is that if several small studies, using the same methodology and hypothesis, are statistically inconclusive, pooling such studies may well lead to a conclusive result. In theory this could be done by merging the original data of the various studies. This would lead to a larger dataset that can be analysed in the same way as the original small studies. In practice, however, it is often impossible to obtain the data of some or all of the previous studies. All that is available are the published summary statistics of the individual studies. An attractive feature of meta-analysis is that a suitable combination (often a weighted average) of these summary statistics may be sufficient to test the hypothesis of interest and then the meta-analyst does not need to have access to the original data of the various studies.

Meta-analysis has become particularly popular in experimental research, for example summarising results from several clinical trials in medical research. More recently, meta-analysis has been increasingly applied to the predominantly non-experimental social sciences, such as economics. Reference works on meta-analysis in general include Cooper and Hedges (1994), Hunt (1999) and Hunter and Smith (2004). In economics, see for example Stanley (2001; 2008) and Florax et al. (2002).

Besides the statistical efficiency gain from pooling estimates (combining inconclusive results may yield conclusive results, as noted above), meta-analysis can also yield significant cost savings. Primary research projects are expensive. If the results of previous research are "transferable" to a new, as yet unexplored, situation, this would avoid the need to conduct a new study. This aspect is particularly important in environmental research in which the economic value is needed of things that are not directly available through market prices, such as the value of lives saved or the value of

public recreational facilities. “Value transfer” is then a particularly useful product of meta-analysis (Brouwer, 2000).

Another benefit of meta-analysis is that by systematically cataloguing the characteristics of the studies conducted to date, it is possible to identify particular combinations of study features that have not yet been explored. Hence meta-analysis can be a systematic tool to design the next empirical study. Later in this chapter it will be shown how, across a large number of studies, the study conclusion can be linked to study characteristics by means of a “meta-regression analysis” (MRA). If the MRA provides a very good model of the range of outcomes observed across studies, it is even possible that the outcome of the next primary study can be predicted by means of the MRA before such a primary study has actually been conducted! In practice, additional primary research is usually worthwhile in any case because circumstances change, sometimes in unobserved ways, rendering the MRA of past research only an imperfect explanation of research findings.

The remainder of this chapter is organised as follows. The next section describes a very simple stylised example of meta-analysis. Section 3 provides the core of the chapter and describes in a step-by-step way how a meta-analysis is conducted. Section 4 briefly outlines how meta-analysis can be conducted when the primary study results are summarised in a categorical way. Section 5 sums up.

2. A simple example

In order to focus on the key aspects of the technique, it is useful to start with a highly simplified, but realistic, example. The example originates from environmental policy analysis. Most governments agree that car exhaust emissions should be reduced, particularly in cities. Such exhaust emissions are a major contributor to the greenhouse gases that have triggered global climate change (leaving aside the debate to what extent climate change has anthropogenic causes). Besides implementation of the various technologies that make cars emit less exhaust fumes, governments may also encourage policies that reduce the total use of private motor vehicles. Not only does this aid countries in meeting international obligations to reduce greenhouse gas emissions, but it also reduces congestion in cities and improves air quality. We would expect that a simple instrument to affect car travel is the fuel tax per litre of fuel. An increase in fuel tax may be expected to reduce vehicle kilometres travelled (VKT) per car, but by how much? This is a typical situation in which there are many studies, but most of the available evidence is observational (i.e., non-experimental): researchers observe how VKT differs across locations (countries, regions, etc.) or points in time and then relate that by means of regression analysis to the various “determinants” of VKT, including the fuel tax charged. There have been many primary studies that use regression models to answer this question. These are reviewed in Hirota and Poot (2005), who themselves estimated a regression model that used observations from 68 cities around the world. Besides many primary studies, there has also been a meta-analysis that informs on the impact of fuel tax on VKT. Espey (1998) did a meta-analysis of the closely related question of the price elasticity of the demand for gasoline.

Basically, the primary research can be expressed by means of the following regression model:

$$\ln(y) = a - b \ln(x) + \text{other factors} + \text{error term} \quad (1)$$

in which ln(y) is the natural logarithm of VKT, ln(x) is the natural logarithm of the fuel tax per litre of fuel, a is a constant and b is the parameter of interest. Because of the use of logarithms, b can be interpreted as follows: a 1% increase in the fuel tax (for example in cents per litre) would decrease VKT by b%. Once data have been collected on y and x, the parameters a and b can be estimated by regression methods, such as Ordinary Least Squares (OLS).

Now assume that three primary studies have been conducted. Study A used a panel of observations from 50 US states observed in 1970, 1980 and 1990. The estimate of b is −0.1 with an estimated standard error se = 0.067. Hence the t-statistic associated with this coefficient (b/se) equals −1.5. The study suggests that an increase in fuel tax of 1% lowers VKT by 0.1%, but the coefficient is not statistically significant, not even at the 10% level (the critical value for that is 1.645). Consequently, we cannot reject the hypothesis that an increase in fuel tax has *no* influence on VKT (and therefore CO_2 emissions) at all.

The second study (B) used 40 annual Australian observations between 1965 and 2005. This study found that b equals −0.4 with a standard error of 0.235. Again, b is not statistically significant (t = −1.7). Finally, Study C uses observations from a cross-section of 68 cities around the world and found a value of b = −0.15 with a standard error of 0.115. Hence t = −1.3. We conclude that in all three studies we cannot reject the hypothesis that a fuel tax (at the levels observed in practice) has no effect on the average number of kilometres that cars travel per year.

What can we conclude when these three studies are combined in a meta-analysis? Intuitively, we might like to simply take the average of the three elasticities. The average value of b across A, B and C is about −0.22. But how reliable is this estimate? Clearly, it makes sense to give more weight to the primary study estimates that have greater precision (smaller standard errors). Statistical theory shows that a weighted average of primary study estimates that uses weights proportional to the inverses of the true variances of the individual estimates has the lowest variance among all linear weighting schemes (e.g. Shadish and Haddock, 1994, p. 265).

When combining regression coefficients from different studies, the simplest assumption that can be made is that there is one "true" value of the parameter of interest and that all studies provide estimates of this parameter. In the literature, this is referred to as the fixed effects (FE) model. Formally, when there are K studies we observe the estimates b_1, b_2, ..., b_K of the parameters β_1, β_2, ..., β_K. In meta-analysis, each estimate is commonly referred to as an "effect size". These effect sizes have estimated variances v_1, v_2, ..., v_K. Under the FE model we assume $\beta_1 = \beta_2 = ... = \beta_K = \beta$, a common effect. Then the weighted average effect size of the K studies is calculated as

$$\bar{b}_{FE} = \sum_{i=1}^{K} w_i b_i \qquad (2)$$

in which the w_i are weights with $w_i = \dfrac{1/v_i}{\sum_{i=1}^{K} 1/v_i}$ and therefore $\sum_{i=1}^{K} w_i = 1$. The weighted average effect size $\bar{b}_{FE}$ has estimated variance $\bar{v}_{FE}$, with

$$\bar{v}_{FE} = \frac{1}{\sum_{i=1}^{K} 1/v_i} \quad (3)$$

The latter can be used to construct a 95 percent confidence interval for the weighted average effect size in the usual way.

Returning to our example, the calculations can be easily carried out by spreadsheet software, such as *Excel*. Much larger and more realistic meta-analyses, which may have hundreds of effect sizes, usually also start with coding the data into a spreadsheet, although the subsequent statistical analysis is often done with specialised software, such as the meta-analysis commands in the statistical software package *Stata* (see Sterne, 2009). Table 1 shows the spreadsheet for our simple example.

Table 1 about here

Using the numbers reported in Table 1 and the equations above, we see that the weighted meta-estimate is 225/318 x –0.1 + 18/318 x –0.4 + 75/318 x –0.15 =–0.129. The standard error is the square root of 1/318 which is 0.056. Because the 95% confidence interval (–0.129 ± 1.96 x 0.056) runs from –0.239 to –0.019 and no longer includes zero, we can conclude that the meta-estimate is now statistically significant, even though the individual studies were not! This is of course just due to the chosen numbers. It is easy to calculate that if just the estimate from study A was changed from –0.1 to –0.02, and everything else remained the same, the 95% confidence interval for the FE meta-estimate would run from -0.049 to 0.003 and the combined effect would therefore be statistically insignificant, just like the individual studies.

Based on the original numbers we conclude that a 10% increase in the fuel tax is likely to reduce vehicle kilometres travelled VKT by 1.29%. Visually, the statistical significance of the FE meta-estimate is reinforced by Figure 1, in which the 95% confidence interval for the regression coefficient *b* crosses the horizontal axis for each of the three studies, but not for the meta-estimate.

Figure 1 about here

There is a big caveat with the meta-analysis described so far: the estimated *b* coefficients are assumed to have been drawn from the *same* distribution. In reality, this is highly unlikely. The other factors in equation (1) are likely to be quite different across studies. From econometric theory we know that the simplest assumption in regression models is that the data in a particular study have been generated by the matrix equation $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with **y** a vector of observations on the dependent variable. The matrix **X** now contains all the variables that matter in the particular primary study. The

vector β represents the coefficients of these variables and ε is a vector of identically and independently normally distributed errors with mean 0 and variance σ^2. In that case, the researcher should estimate β by OLS: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, which is normally distributed with mean β and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. So even in the simplest case where β is the same for each study but the data differ between studies, the estimated effect would have a different variance in each study, which depends on the covariates. This is referred to as *heteroscedasticity* among the pooled estimates. Moreover, when **X** and **y** are very different across studies, it is plausible that the true parameter vector β varies across studies too. This is referred to as *heterogeneity*. Moreover, it may well be that some researchers estimated the "wrong" model, that is a model that is not the correct representation of the underlying data generating process. This is referred to as *misspecification*, which can make the estimated **b** biased (when relevant variables have been omitted) or inefficient (when irrelevant variables have been added); see for example Gujarati and Porter (2009). Naturally, specification errors will have an impact on pooled estimates such as the FE meta-estimate. Finally, the meta-analyst may not actually observe all studies that have been conducted, but only the ones that provide strong evidence for the issue under consideration. This is referred to as *publication bias*. It is clear that in the case of heteroscedasticity, heterogeneity, misspecification bias and publication bias, the procedure we carried out for the simple example is not statistically appropriate. Most researchers estimate in this more realistic, but complex, case a meta-regression model in which estimates from various studies are explained by a range of study characteristics. The appropriate specification of the model for MRA is still subject to some debate (see e.g. Becker and Wu, 2007; Koetse et al. 2010). We will return to MRA in the next section.

To sum up the discussion to this point: a meta-analysis "pools" results where you can't "pool" the data. The meta-estimate tends to be a weighted average of the estimates from primary studies. It gives more weight to more precise estimates. In some cases, as in the example, the meta-analysis may be conclusive when the individual studies are not. The approach used here assumes that the meaning of *b* is the same in all studies (that is satisfied in the example, where each *b* is an elasticity) and that there is one "true" value of *b* around the world. However, the estimates are derived in the social sciences from observational data and unlikely to come from controlled experiments such as randomized trials. This necessitates checking for "moderator variables" (study characteristics that may affect the statistical results). The next section discusses the process of meta-(regression) analysis on a step-by-step basis.

3. Meta-analysis in 10 "easy" steps

To keep this chapter of reasonable length, much detail must be omitted, but there are already many excellent detailed descriptions of how to do meta-analysis, such as in the references in the introductory section. The procedure is summarised in Table 2 in 10 steps. Step 1 of meta-analysis consists of the collection of large and representative sample of all empirical studies on a particular topic. Each study must report a quantitative analysis of "real world" data. Studies are nowadays easily found with search engines such as *EBSCO Host*, *Google Scholar*, *EconLit*, *RePEc*, etc. An alternative method is to use the so-called snowballing procedure: start with a few most commonly

cited studies, then check the references of each of these studies for additional articles of interest, and so on.

Table 2 about here

Step 2 is to choose an effect size that can be observed in each study: a parameter estimate, *t* statistic, correlation coefficient, etc. Of course, parameter estimates can vary with units of measurement. Hence, unit free measurement such as elasticities or beta coefficients (regression coefficient times the standard deviation of the independent variable divided by the standard deviation of the dependent variable) are preferred. If such unit free measurement is not possible, or if the variables of interest are defined rather differently across studies, meta-analysts can simply use the *t* statistics linked to the variable of interest in the primary study and convert *t* statistics into partial correlation coefficients

$$r_i = \frac{t_i}{\sqrt{t_i^2 + N_i - K_i}} \text{ with } \operatorname{var} r_i = \frac{\left(1 - r_i^2\right)^2}{N_i - 2} \tag{4}$$

in which t_i is the *t* statistic of the coefficient of interest in the i^{th} study, N_i the number of observations in study *i* and K_i the total number of explanatory variables in study *i*. An alternative is Fisher's *Z* value derived from this partial correlation coefficient:

$$Z_i = 0.5 \times \ln\left[\frac{1 + r_i}{1 - r_i}\right] \text{ with } \operatorname{var} Z_i = \frac{1}{N_i - 3} \tag{5}$$

Sometimes meta-analysts simply summarise the results of a study in terms of a categorical variable: a study may conclude that the effect is statistically significant and negative, or statistically insignificant, or statistically significant and positive (see de Groot *et al.* 2009 and Card et al. 2010 for recent examples). Like the regression coefficients themselves, the pooling of measures of statistical significance will also be affected by the covariates of the primary studies. In practice, this is taken into account by incorporating many distinguishing features of the primary studies in the MRA. If a primary study estimated a "wrong" model (by omitting relevant variables or including irrelevant ones), that naturally also affects the statistical inference when pooling the effect sizes (Keef and Roberts, 2004). Again, this problem is in practice addressed by specification and estimation in MRA.

Step 3 of meta-analysis consists of entering the effect sizes and relevant study characteristics into a spreadsheet. This coding of study information is the most time consuming task of meta-analysis. In advance it is not always clear which study characteristics matter and whether this information is available for all collected studies. This implies that the coding is often an iterative process. This

process should not be purely data driven, but also be strongly guided by theory. The meta-analyst should have a good understanding of the field and what kind of variables and issues are likely to affect the causal relationship he or she is interested in. Because there are sometimes also differences in opinion about interpretation of information from any given study, the coding will be ideally checked by one or more co-authors of the meta-analysis.

Another issue in Step 3 is that each individual study often reports many estimates. In that case, one could either just select the estimate preferred by the author(s), or code all estimates. Because the range of estimates that result from a sensitivity analysis within each study is usually informative about the robustness of the findings, it is preferable to include as many estimates as possible. Nonetheless, it is clear that pooling several estimates from a particular study will have different statistical implications than pooling estimates across studies. This is taken into account in more advanced MRA techniques.

Several of the study characteristics to be coded are numbers: the number of observations used in the primary study, the final year of the data, the number of geographical units, etc. Other study characteristics are qualitative variables with various levels. For example, some researchers may use cross-sectional data, while others use time-series data. Alternatively, such data could have been combined in the form of a panel (pooled cross-section time-series data). Qualitative information is recorded by means of dummy variables. For example, a dummy variable "*ts*" could be assigned the value 1 when an effect size was obtained by time-series analysis and the value 0 otherwise.

The fourth step consists of an exploratory and descriptive analysis of the coded information. This can be done in spreadsheet programs like *Excel* or in statistical software such as *SPSS* and *Stata*. One of the simplest statistics that can be calculated is the percentage of effect sizes that confirms the stated hypothesis. This procedure is referred to as "vote counting". This percentage can be calculated for different groups of studies and the results presented in tabular form. It has been shown that vote counting on its own is best seen as just a purely descriptive device, although it is possible to use certain regression models, such as the ordered probit model, to explain in what circumstances a primary study is likely to yield a statistically significant result. This will be elaborated in the next section.

One of the common criticisms of meta-analysis is that studies are often so different that we are "comparing apples and pears": there is not a single underlying "true" parameter. This is formally tested in Step 5 of meta-analysis by the *homogeneity* test. A test of the hypothesis that studies do in fact share a common true effect size uses the following homogeneity statistic:

$$Q = \sum_{i=1}^{K} \left[\left(b_i - \bar{b}_{FE} \right)^2 / \left(v_i \right) \right] \tag{6}$$

If Q exceeds the upper-tail critical value of the chi-square distribution with $K-1$ degrees of freedom, the observed variance in effect sizes is considerably greater than what we would expect by chance if all studies shared the same "true" effect size. It can be easily checked that in the example of the previous section, Q = 1.549, which should be compared with the 5% significance critical value

of the chi-square distribution with 2 degrees of freedom, which is 5.991. We can therefore accept the hypothesis that the three studies in the example were sufficiently similar for the sample of studies to be considered statistically homogeneous and the FE estimate appropriate. When within-study sample sizes are large (and estimated variances of the effect sizes small), homogeneity is likely to be rejected even when the individual effect sizes do not differ much. Moreover, homogeneity is also likely to be rejected we have a large sample of non-experimental primary studies in the data set of effect sizes. The best way to account for heterogeneity is then to use regression techniques (see Step 7 below).

An intermediate approach, between the FE approach and the regression approach, is to assume that the underlying parameter differs between studies, but is drawn randomly for each study from a particular distribution. This is referred to in the literature as the random effects (RE) model. In this case, the "true" elasticity β_i of study *i* is assumed to be distributed with mean β and variance $v_i^* = \sigma_\beta^2 + v_i$, where σ_β^2 represents the between-studies variance and v_i represents the within-study variance. It can be shown (e.g., Shadish and Haddock 1994, p.274) that an unbiased estimate of σ_β^2 is given by

$$\hat{\sigma}_\beta^2 = \left[\sum_{i=1}^{K} b_i^2 - \left(\sum_{i=1}^{K} b_i\right)^2 / K\right] / (K-1) - \left(1/K\right)\sum_{i=1}^{K} v_i \qquad (7)$$

and estimates of v_i^* are therefore calculated as $v_i^* = \hat{\sigma}_\beta^2 + v_i$. The weighted mean elasticity and its estimated variance can then be computed by replacing v_i by v_i^* in equations (2) and (3). Hence the weighted average effect size of the *K* studies is calculated as

$$\bar{b}_{RE} = \sum_{i=1}^{K} w_i^* b_i \qquad (8)$$

in which w_i^* are weights with $w_i^* = \dfrac{1/v_i^*}{\sum_{i=1}^{K} 1/v_i^*}$ and therefore $\sum_{i=1}^{K} w_i^* = 1$. The weighted average effect size $\bar{b}_{RE}$ has estimated variance $\bar{v}_{RE}$, with

$$\bar{v}_{RE} = \frac{1}{\sum_{i=1}^{K} 1/v_i^*} \qquad (9)$$

After calculating the new variances, exactly the same calculations can be done as in the example in Table 1. The reader can check that in that simple example $\hat{\sigma}_\beta^2$ = 0.00146 . The RE meta-estimate is -0.134, with a 95% confidence interval running from -0.257 to -0.011. The RE meta-estimate tends to be closer to the ordinary mean of the individual effect sizes than the FE meta-estimate. Moreover, the confidence interval based on the RE meta-estimate is wider than that of the FE meta-estimate.

While in the simple example discussed in this chapter, the *Q* statistic suggested that the three studies are homogeneous, in many applications in the social sciences this hypothesis will be rejected. In that case, the meta-analyst needs to find so-called *moderator* variables that can explain the "excess variation" in study outcomes by means of a regression model. This is Step 6 of the meta-analysis. Such moderator variables must be carefully chosen because we can get a biased explanation of differences in study outcomes when, in the regression analysis, omitted study characteristics that do matter are correlated with the included study characteristics. This omitted variables bias can be reduced by taking into account as many study characteristics as are theoretically plausible. On the other hand, a strong correlation among some of the moderator variables themselves leads to the well-known problem of *multicollinearity* in regression analysis and this should be avoided.

Step 7 of meta-analysis consists of running a meta-regression analysis (MRA) in which the effect sizes are explained in terms of the moderator variables. Again let's assume that there are *K* studies and we observe the effect sizes b_1, b_2, ..., b_K that correspond to the "true" parameters β_1, β_2, ..., β_K. These effect sizes have estimated variances v_1, v_2, ..., v_K. Now we assume that there are *P* known moderator variables M_1, M_2, …,M_P that are related to the effect sizes via a linear model as follows:

$$b_i = \beta_i + \eta_i = \gamma_0 + \gamma_1 M_{i1} + \gamma_2 M_{i2} + ... + \gamma_P M_{iP} + \eta_i \qquad (10)$$

in which M_{ij} is the value of the *j*th moderator variable associated with effect size *i* and the η_i is the disturbance term. Because the effect sizes are heteroscedastic, this regression model should not be estimated by OLS. The simplest estimator to use is the Weighted Least Squares (WLS) estimator, which is included in all statistical software packages. The weights variable that is specified in the software command is the vector of reciprocals of the variances of the primary study estimates ($1/v_i$). In this case, WLS is equivalent to running OLS on transformed data in which each row of primary study data is divided by the standard error of the effect size $\sqrt{v_i}$ (see, e.g. Gujarati and Porter, 2009). However, standard regression packages do not estimate the standard errors of the regression coefficients of the moderator variables in the MRA correctly. A correction can be made (see e.g. Hedges 1994, p.296). Alternatively, specialised software for meta-regression analysis can be used. For example, it may be argued that "best practice" meta-regression model in the absence of publication bias is the combination of (10) with the RE model of equations (7) to (9) (see Harbord and Higgins, 2008). This model, sometimes referred to as the Mixed Effects (ME) model can be estimated by the command *metareg* in *Stata* (see Sterne, 2009). Nonetheless, Koetse et al. (2010) demonstrate that in certain situations that are common in non-experimental social sciences, WLS may be preferred (which can be estimated with the command *vwls* in *Stata*), but they also note that

for large samples both methods perform well. Consequently, it pays to collect as many effect sizes as possible before doing a MRA.

When using the WLS or ME meta-regression models, the weights variable $1/v_i$ may be adjusted for two reasons. One is that if some primary studies only report one estimate, while others report many estimates, the studies reporting many estimates may get too much weight in the MRA. The simplest solution to that is to multiply $1/v_i$ by $1/k$ where k is the number of effect sizes that come from the study that included observation i. In practice, it pays to also experiment with multiplying $1/v_i$ by $1/m$ where $1 \leq m \leq k$. A second reason for adjusting the weights variable $1/v_i$ is variation in the quality of the estimates. In that case, we can multiply $1/v_i$ by q, where q is a "score" of quality, such as the impact factor of the journal in which the result was published. As this is rather arbitrary, an alternative approach to account for quality is to introduce dummy variables among the moderator variables that represent particular types of outlets (top journal, average journal, working paper, etc.). In that way, we can test whether top quality journals generate on average larger or smaller effect sizes.

In Step 8 of meta-analysis, we test for publication bias and correct for it if it has been found to affect the available data. As noted earlier, publication bias can arise if studies that are convincing are more likely to be published than studies in which the estimated coefficient of interest turned out to be statistically insignificant. In our simple example, each of the three studies A, B and C found a statistically insignificant effect. It may have been hard to get such studies published in good journals because editors and referees prefer studies that confirm the hypothesis that such studies set out to test. If a researcher gets an inconclusive result, he or she may not even bother submitting it to a journal and, instead, simply file the draft paper away. This explains why publication bias is also referred to as "file drawer bias". To avoid file drawer bias, the meta-analyst should include unpublished working papers in the sample of studies. These are now often downloadable, but it is sometimes useful to write to researchers who have worked on a particular topic to ask them if there are any estimates that have not been made public but that can be released for the meta-analysis.

A very popular procedure to test for publication bias is the so-called funnel plot, which is a simple scatter diagram in which the reciprocal of the estimated standard error (i.e., $1/\sqrt{v_i}$), is plotted on the vertical axis and the corresponding effect size on the horizontal axis. The former is usually referred to as the *precision* of the effect size. A detailed discussion can be found in Stanley and Doucouliagos (2010). Figure 2 displays an example from Nijkamp and Poot (2005). The effect sizes in Figure 2 refer to the so-called *wage curve*: the elasticity of the relationship between the wages individuals receive and the unemployment rate in their local labour market. The FE meta-estimate for this dataset is about -0.06. This can be interpreted to say that a 10% increase in the unemployment rate (from, say, 5% to 5.5% of the labour force) would lower wages by about 0.6%. Blanchflower and Oswald (1994) thought that there would be, roughly, one "true" universal underlying elasticity that should apply to all studies around the world and they estimated this elasticity to be about -0.1. In fact, the simple average of the effect sizes depicted in Figure 2 is indeed -0.1, but it was noted in Section 2 that the FE estimate is statistically preferred to the ordinary average. In any case, if there was a true underlying value common to all studies, then the funnel plot should be symmetric around that value and look like a funnel because the most precise estimates (usually the ones from primary studies with many observations) would cluster closely to the "true" value. However, the funnel plot from

the wage curve literature shows some bias: there are very few primary studies that show positive values even when the precision is low (i.e., small values of $1/\sqrt{v_i}$). Researchers who found such positive elasticities may have discarded their computer output because they thought that their results were "odd" and not publishable given that after the publication of Blanchflower and Oswald (1994) researchers started to expect an estimate of around -0.1.

Figure 2 about here

However, while publication bias will lead to an asymmetric funnel plot, funnel plots can also be asymmetric when there is no publication bias. For example, if there is substantial heterogeneity then there is not just one "true effect" but several and only the funnel plots of homogeneous sub-samples of effect sizes would look symmetric, not a funnel plot that plots all points together. In the presence of heterogeneity, it is possible to check for publication bias after an MRA has been conducted. In that case the funnel plot would be a plot of precision of the primary study estimates against the *residuals* of the meta-regression model. This funnel plot should be roughly symmetric around the vertical axis.

Besides the funnel plot, there are various other ways to detect publication bias. One simple idea is that, in primary data analysis, researchers often run a range of regression models and only report those specifications in which the estimated coefficient of the variable of interest has a *t*-statistic larger than the "magic number" 1.96, or roughly 2 (which would imply that the chance – usually referred to as the *p* value – that the specific results are obtained while the true parameter is zero, is less than 0.05). On the other hand, researchers do sometimes obtain *t* statistics that are smaller than 2 even though there is a true effect. There is likely to be publication bias when the sample of effect sizes will contain relatively few *t* statistics that are less than 2. By definition, the *t* statistic is the estimated coefficient divided by its standard error. If these bunch around 2, the reported effect sizes will be proportional to the corresponding estimated standard errors. If the primary studies were obtained from the *same* data generating process, the estimated effect sizes b_i and their estimated standard errors $\sqrt{v_i}$ should be uncorrelated. Formally, publication bias arises when $\delta \neq 0$ in the equation $b_i = \beta + \delta\sqrt{v_i} + \varepsilon_i$, with β the true effect as before and ε_i an error terms that accounts for the differences between studies (Card and Krueger, 1995). If we divided both sides of this equation by $\sqrt{v_i}$ we get:

$$t_i = \beta\,(1/\sqrt{v_i}) + \delta + \xi_i \qquad (11)$$

Stanley (2008) shows that this equation can be estimated by OLS. The estimates can be used to test for publication bias ($\delta \neq 0$) and for the presence of a genuine effect ($\beta \neq 0$) in the usual way. The former test is called the funnel asymmetry test (FAT), the latter the precision effect test (PET).

However, in MRA there will be usually heterogeneity in which β varies across studies due to differences in study characteristics. In that case, the regression model that accounts for publication bias is usually specified as

$$b_i = \gamma_0 + \gamma_1 M_{i1} + \gamma_2 M_{i2} + ... + \gamma_P M_{iP} + \delta(1/\sqrt{v_i}) + \omega_i \quad (12)$$

This is the same as (10), but now with the "publication bias correction" regressor $1/\sqrt{v_i}$ added. Feld and Heckemeyer (2011) review various ways to estimate this model.

Another way to deal with publication bias is to model explicitly the phenomenon that researchers are likely to discard regressions with small *t* statistics (which imply high *p* values) and report all regressions with large *t* statistics (small *p* values). Hedges (1992) formulated a statistical model that estimates the probabilities that certain regressions will not be reported. He estimated this model by means of a maximum likelihood method. Combined with MRA, this idea has been applied to a range of topics (see Stanley, 2008), including the wage curve research that coincides with the funnel plot in Figure 2. In MRA, the methodology has not always led to meaningful results, but Nijkamp and Poot (2005) found that, in their MRA of the wage curve literature, the probability that studies with *p* values of less than 0.01 were published could be assumed to be 100%. In contrast, studies with *p* values of between 0.01 and 0.05 were only reported with an estimated 53% chance, while those with *p* values greater than 0.05 were only reported with 29% probability. Taking these results literally, they suggest that almost three quarters of findings from estimating wage curves that had *t* statistics less than about 2 on the coefficient of the unemployment rate, remained unpublished!

Because this method to control for publication bias is not yet included in standard statistical software, readers may prefer to focus on estimating equation (12), for example with the *metareg* command in Stata. One informal way of testing for publication bias is to see if results that have been reported in refereed journal articles are different from those that were reported in unpublished working papers. It could be argued that in today's "publish or perish" research environment in which it is very easy to post results in working paper form on the internet, the likelihood of "file drawer bias" is diminishing. Publication bias can then simply be tested by including in the MRA a dummy variable *pb* that takes the value 1 for an effect size that was published in a refereed journal and 0 otherwise. If the coefficient of *pb* is statistically insignificant (i.e. the hypothesis that *pb*=0 cannot be rejected) we can conclude that there is no publication bias of this specific type (the alternative interpretation is that the quality of the outlet (refereed/working paper) does not matter). Of course, there remains even then still the possibility that some studies, published or unpublished, reported only results that were consistent with the researcher's prior beliefs. In any primary study, the sample of data, the selected variables and the statistical model should be varied to see whether the coefficient of interest in a regression model is robust to such specification choices. This is referred to as *sensitivity analysis*. Although space usually limits the amount of sensitivity analysis that can be reported in any given paper, a fair and frank assessment should be included regarding the robustness of the findings of the primary study. If the findings are not robust, further secondary analysis may be needed at a later stage.

Sensitivity analysis is also sound practice in meta-analysis. This is Step 9. It would consist of varying the sample of effect sizes, the MRA model and the moderator variables. Included in Step 9 would be a re-run of the homogeneity test (6), but now on the residuals of the meta-regression model rather than the effect sizes themselves. Clearly, with appropriately selected moderator variables, the *Q* statistic on the residuals will become much smaller than on the original effect sizes. In *Stata*, the *metareg* command will actually report the proportion of between-study variance that is explained by

the moderator variables. This is equivalent to the (adjusted) R^2 in the ordinary regression model. In sensitivity analysis, a desirable and robust specification would be the MRA with the moderator variables selected such that it has – relative to other specifications with the same sample – a high adjusted R^2, even when holding some of the meta-observations back.

Writing up a summary of such sensitivity analysis can be included in the final step of meta-analysis, Step 10. In this step, the entire project is written up and submitted for publication. If sensitivity analysis shows that the results are strongly sensitive to specific choices and "strange" results are not found to be due to programming or data errors that can be corrected, then the meta-analyst should resist the temptation to omit such unusual results from the write up. Failing to fully disclose the results from sensitivity analysis may render the meta-analysis just as sensitive to publication bias as the primary studies on which it is based! However, as in primary research, there is usually limited scope to include a detailed discussion of sensitivity analysis in the published paper. Sometimes additional results can be posted on a website.

4. Meta-analysis of categorical findings

In order to obtain a stylised fact on a particular topic in the spatial sciences, for example the extent to which changes in the average house price in a city spill over to house prices in surrounding cities within a 100 km radius, it would be helpful to have the research replicated to different parts of the country or to different time periods, or to different countries. However, pure replication is often looked down upon by researchers in the social or spatial sciences because it would not be considered sufficiently innovative to yield a publication in a highly ranking journal (Hamermesh, 2007). Instead, researchers will aim to introduce new ways to measure the data, new estimation techniques, new specifications, etc. But while such innovative activity is laudable and enhances the stock of knowledge, it complicates meta-analysis through generating extensive heterogeneity, not only in terms of moderator variables and functional forms, but also in the measurement and interpretation of the effect sizes themselves.

In some cases estimates can be made comparable. For example, a Japanese study may show that an increase in the price of a Tokyo to Osaka Shinkansen (bullet train) ticket by 1000 yen reduces daily passenger numbers by 3000, whereas a similar study of the French TGV expresses the relationship in terms of a demand elasticity and finds, for example, an elasticity of -0.3, i.e. a 1% in the ticket price reduced passenger numbers by 0.3%. If we know the average ticket price in Japan and the average daily number of passengers, the two studies can be made comparable by converting the Japanese finding into an elasticity also. Because elasticities are dimensionless, they are very useful for comparisons across studies.

However, often insufficient information is available to calculate elasticities. Alternatively, the measures used in various studies are conceptually different and cannot be compared at all. For example, in studies of the impact of competition within industries and of diversity among industries on productivity growth of firms, there are various ways in which "competition" and "diversity" can be measured and these measures are not always directly comparable (e.g., de Groot et al. 2009). In that case, there are several ways in which the results of such studies can still be combined in order

to draw some general conclusions. In this section three methods of that type will be discussed. They either use information on statistical significance only or draw conclusions of a categorical nature. They are: an MRA of Fisher's *Z* values; the use of ordered probit models; and the use of rough set analysis.

A focus on statistical significance of study results was already briefly discussed in Section 3 where the partial correlation coefficient r_i and Fisher's *Z* value Z_i were defined in terms of the reported *t* statistics. In principle, MRAs can be formulated with, for example, Fisher's *Z* value on the left hand side and a range of moderator variables in the right hand side. Substituting (5) on the left hand side of (10), we get

$$Z_i = \gamma_0 + \gamma_1 M_{i1} + \gamma_2 M_{i2} + ... + \gamma_P M_{iP} + \tau_i \qquad (13)$$

with τ_i the corresponding error term. This model can be estimated by Weighted Least Squares, with the weights variable being the vector of N_i's. This kind of MRA will assist in identifying which kind of study characteristic leads to strongly positive or negative *Z* values.

However, high statistical significance does not necessarily imply high economic significance. Yet the latter may be more important from the policy perspective. For example: it is not surprising and not very useful to know that all studies that look at the determinants of a person's earnings find that years of schooling is statistically significant at the 0.1% level. It is far more useful to know by how much earnings goes up with an additional year of education (also called the rate of return to an additional year of schooling). The former conclusion refers to statistical significance, the latter to economic significance. MRAs of partial correlation coefficients or *Z* values only explain statistical significance, they do not inform on economic significance.

In any case, even if only statistical significance can be compared, there is an alternative method to use, instead of model (13). This alternative has become popular in recent years. In this alternative method, the study outcomes are recorded as follows: significantly negative, insignificant (negative or positive), or significantly positive. Hence the effect size is now a categorical variable that can take on three levels. The probability that any study takes on one of these three levels can be assumed to be a function of a set of study characteristics. The statistical model that relates these probabilities to explanatory variables is called the *ordered probit* model. Examples of this approach are Card et al. (2010) and de Groot et al. (2009) (with the latter distinguishing between insignificantly negative and insignificantly positive as two separate categories rather than one category).

It should be noted that by recording study outcomes as significantly negative, insignificant or significantly positive, the information on the extent of statistical significance which is contained in, for example, the *t*-statistics or *p* values of the estimated coefficients, is ignored. This can be an advantage if one does not want to give too much weight to primary study results obtained from very large data sets (such as population census counts), which are likely to generate very large *t* statistics, as compared with primary study results obtained from well-designed sample surveys with fewer observations but a large range of variables.

The ordered probit model in meta-analysis assumes the presence of a latent variable, y^*, which can be interpreted as the unmeasured degree of "conclusiveness" of the study. It is assumed that y^* can be explained by a set of moderator variables M_i as follows:

$$y^* = \sum_{i=1}^{P} \theta_i M_i + \psi \tag{14}$$

where ψ is an error term that is assumed to identically and independently normally distributed with mean 0. What we actually observe is information on the categorical variable y which coincides with the three categories discussed above; with y = 0 implying that the coefficient of interest in the primary study is significantly negative, y = 1 insignificant, or y = 2 significantly positive. This observed variable has the following structure:

$$\begin{aligned} y &= 0 \quad if \quad y^* \le \mu_1 \\ y &= 1 \quad if \quad \mu_1 < y^* \le \mu_2 \\ y &= 3 \quad if \quad y^* > \mu_2 \end{aligned} \tag{15}$$

The μ-parameters, along with the θ's, are estimated by the maximum likelihood estimator of the ordered probit model. It is important to note that the interpretation of the estimated coefficients of an ordered probit analysis requires some care, see e.g. Verbeek (2004, Chapter 7).

The final method to be discussed in this section is useful in the cases where the methodologies adopted by the primary studies vary widely. In that case it would be difficult, if not impossible, to formulate a common statistical model. Nijkamp and Poot (2004) provide an example where the research question of the meta-analysis was to establish whether fiscal policies have an impact on long-run economic growth and, if so, what kind of fiscal policies would be most effective to promote growth. In this literature a wide range of methodologies has been used. However, in each case it was possible to assign a prior belief (expected impact) to a particular policy, such as: "increasing the share of general government spending as a percentage of GDP lowers growth", or "increasing expenditure on education increases growth". Studies can then be tabulated in terms of whether the evidence supported the expected impact or whether the study was inconclusive. Table 3 summarises the conclusions of 123 studies on the impact of fiscal policies on growth. Note that the proportions in the two columns do not always add to 1, because some studies may conclude the *opposite* of what is expected. For example, in defence studies 5% concluded that defence spending increases economic growth.

Table 3 about here

As in MRA, the question again arises to what extent such study outcomes are related to study characteristics. Given the absence of a common statistical model for the studies summarised in Table 3, an alternative method that is available is based on so-called *rough set theory*, which was developed by Pawlak (1982). A summary of the theory and applications to meta-analysis can be found in van den Bergh et al. (1997). Nijkamp and Poot (2004) apply this methodology to the study findings summarised in Table 3.

A *rough set* is a set for which it is uncertain in advance which objects belong precisely to that set, although it is in principle possible to identify all objects that may belong to the set at hand. In meta-analysis, each object represents one primary study. In order to identify objects, we use study characteristics that are referred to in rough set analysis as *attributes*. The spreadsheet of study attributes (including the conclusions drawn) is referred to as the *information matrix*. Meta-observations that have the same values for a given sub-set of attributes are called *indiscernible*. The information matrix can be partitioned into elementary sets of indiscernible objects. Of course, the more attributes are taken into account, the larger the number of elementary sets. The key question is: how many attributes are needed to classify the objects "reasonably well"?

Hence, as in MRA, the objective is to find out which study characteristics matter and which are redundant in relation to study outcomes. Moreover, as in MRA, a sensitivity analysis can also be conducted in which the definition of the variables or the number of meta-observations is varied. However, unlike MRA, all explanatory variables in rough set analysis must be of a categorical nature. So if sample size is an integer variable in MRA, rough set analysis would require a discretisation of this information. For example, sample size could be categorised as follows: "small-size sample", "medium-size sample" or "large-size sample" (referred to as *classes*). Clearly, the proper demarcation of class boundaries requires some skill. As the results may be sensitive to the mapping used, some experimentation is often necessary.

Rough set analysis uses computer software that helps to find those attributes (study characteristics) that are redundant in explaining the objects (studies). The equivalent in MRA is that the regression model tells us which study characteristics are statistically insignificant in explaining the study outcomes. In the rough set analysis of the fiscal impact studies it was found that 8 of 9 attributes were needed to classify the studies (see Nijkamp and Poot, 2004). The only redundant attribute was whether the primary study focussed on the national impact of fiscal policy on growth, or the regional impact. The fact that this aspect did not matter in linking study characteristics with study conclusions is of course an interesting conclusion in itself. It suggested that the results of the meta-analysis are equally informative for designing economic growth policy at the regional level as at the national level.

A rough set analysis will generate a set of *deterministic rules* and will measure the fraction of objects/studies that are completely in accordance with these rules. This fraction is referred to as the relative strength. For example, Nijkamp and Poot (2004) found that "in studies on public infrastructure, using techniques for time series analysis, the impact of infrastructure policy on growth is significantly positive." The relative strength of this statement was 25.5%, i.e. in 10 out of the 39 studies on the impact of infrastructure on growth the previous statement was confirmed.

In conclusion, rough set analysis provides a means for systematically digesting information from a range of very different studies. A criticism of this approach is that it is a form of "pattern recognition" that falls under the realm of artificial intelligence in computer science and is therefore divorced from the statistical foundations that underpin most spatial and social science research. As long as all primary studies used regression analysis, which remains the most commonly used tool in the non-experimental social sciences, it would be preferable to use either the MRA or the ordered probit approaches.

5. Conclusions

This chapter has outlined the art and science of meta-analysis, a quantitative approach to digesting previous empirical research. Meta-analysis can be either the start of a new primary study (replacing or supplementing the narrative review) or become the main focus of the research. Meta-analysis is applicable to both experimental and non-experimental contexts; but each has developed their own techniques. This chapter focussed on those techniques most commonly applied by economists and other social and spatial scientists.

As in primary research, different techniques are often possible for a particular application. In that case, it pays to vary techniques (unless a particular technique is preferred for theoretical reasons) and look for robust results across techniques. “Good” meta-analysis should account for observed heterogeneity (with moderator variables signalling differences in data, specifications, covariates, etc.) and unobserved heterogeneity. It should also test for publication bias, quality differences between studies, and account for the statistical consequences of combining several estimates coming from a single study with estimates coming from different studies.

Philosophically, it could be argued that meta-analysis in the social sciences consists of collecting quantitative "opinions" obtained from a non-randomly selected sample of unique studies and that our ultimate goal is to explain the distribution of these "opinions". This distribution may in fact be just as interesting, or even more interesting, than a “central value” such as a mean effect size.

Nonetheless, readers will often remain interested in the “bottom line”: a “stylised fact” or a value to transfer to another context that is obtained from pooling primary studies. This situation is similar to that of forecasting in which the weighted average of the forecasts of a panel of experts usually outperforms any individual forecast (e.g. Newbold and Bos, 1994). In such forecasting, we actually prefer the methods used by the various forecasters to be very different. Weights in that context are based on past “out of sample” performance (mean squared prediction errors). In the context of meta-analysis, the weights we assign to primary study effect sizes can be based on the reported precision, possibly adjusted for quality variation etc., as outlined previously. As in forecasting, it may pay in meta-analysis to initially hold back some data from the MRA, and then conduct the MRA with the full sample in order to see how sensitive the results are to the exclusion of certain studies.

The methodology for meta-analysis is still evolving. The present chapter has given an outline of current practice in non-experimental social science research. While space was too limited to include full details on the various techniques, it should be possible for the reader to conduct a meta-analysis through following the 10 steps discussed in this chapter. There will be no shortage of research areas to which this newly acquires skill can be applied. In fact, given the “flood of numbers” on human behaviour in the 21st century, and the associated explosion of applied empirical research, the scope for meta-analysis is unlimited!

References

Angrist, J.D and Pischke, J.-S. (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24(2): 3-30.

Becker, B.J. and Wu, M.-J. (2007) The synthesis of regression slopes in meta-analysis. *Statistical Science* 22(3): 414-429.

Blanchflower, D.G. and Oswald, A.J. (1994) *The Wage Curve*. Cambridge MA: MIT Press.

Brouwer, R. (2000) Environmental value transfer: state of the art and future prospects. *Ecological Economics* 32(1): 137-152.

Card, D. and Krueger, A.B. (1995) Time-series minimum-wage studies: a meta-analysis. *American Economic Review* 85: 238-243.

Card, D., Kluve, J. and Weber, A. (2010) Active labour market policy evaluations: a meta-analysis. *The Economic Journal* 120(November): F452-F477.

Cooper, H. and Hedges, L.V. (1994) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

de Groot, H.L.F., Poot, J. and Smit, M.J. (2009) Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis. In: Capello, R. and Nijkamp, P. (eds) *Handbook of Regional Growth and Development Theories.* Cheltenham UK: Edward Elgar, 256-281.

Espey, M. (1998) Gasoline demand revisited: An international meta-analysis of elasticities. *Energy Economics* 20(3): 273-295.

Feld, L.P. and Heckemeyer, J.H. (2011) FDI and taxation: a meta-study. *Journal of Economic Surveys*. Forthcoming.

Florax, R.J.G.M., Nijkamp, P. and Willis, K.G. (eds) (2002) *Comparative Environmental Economic Assessment*, Cheltenham UK: Edward Elgar.

Glass, G.V. (1976) Primary, secondary and meta-analysis of research. *Educational Researcher* 5: 3-8.

Gujarati, D.N. and Porter, D.C. (2009) *Basic Econometrics*. 5th Edition. McGraw-Hill Irwin.

Hamermesh, D.S. (2007) Replication in economics. *NBER Working Paper 13026*. Cambridge MA: National Bureau of Economic Research.

Harbord, R.M. and Higgins, J.P.T. (2008) Meta-Regression in Stata. *The Stata Journal* 8(4): 493-519.

Hedges, L.V. (1992) Modelling publication selection effects in meta-analysis. *Statistical Science* 7: 246-255.

Hedges, L.V. (1994) Fixed effects model. In: Cooper H. and L.V. Hedges (eds) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 285-299.

Hirota, K. and Poot, J. (2005) Taxes and the environmental impact of private car use: evidence from 68 cities. In: A. Reggiani and L. Schintler (eds) *Methods and Models in Transport and Telecommunications: Cross-Atlantic Perspectives*. Berlin: Springer Verlag, 299-317.

Hunt, M.M. (1999) *How Science Takes Stock: The Story of Meta-Analysis*. New York: Russell Sage Foundation.

Hunter, J.E. and Schmidt, F.L. (2004) *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. New York: Russell Sage Foundation, 2nd ed.

Keef, S.P. and Roberts, L.A. (2004) The meta-analysis of partial effect sizes. *British Journal of Mathematical and Statistical Psychology* 57: 97-129.

Koetse, M.J., Florax, R.J.G.M. and de Groot H.L.F. (2010) Consequences of effect size heterogeneity for meta-analysis: a Monte Carlo study. *Statistical Methods and Applications* 19(2): 217-236.

Leamer, E. (1983) Let's take the con out of econometrics. *American Economic Review* 73(1): 31-43.

Newbold, P. and Bos, T. (1994) *Introductory Business & Economic Forecasting*. Cincinatti, Ohio: South-Western Publishing.

Nijkamp P and Poot J (2004) Meta-analysis of the impact of fiscal policies on long-run growth. *European Journal of Political Economy* 20(1): 91-124.

Nijkamp, P. and Poot, J. (2005) The last word on the wage curve? *Journal of Economic Surveys* 19(3): 421-450.

Pawlak, Z. (1982) Rough Sets. *International Journal of Information and Computer Science* 11, 341-356.

Pearson, K. (1904) Report on certain enteric fever inoculation statistics. *British Medical Journal* 3: 1243-1246.

Shadish, W.R. and Haddock, C.K. (1994) Combining estimates of effect size. In: Cooper H. and L.V. Hedges (eds) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 261-281.

Stanley, T.D. (2001) Wheat from chaff: meta-analysis as quantitative literature review. *Journal of Economic Perspectives* 15: 131-150.

Stanley, T.D. (2008) Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics* 70(1): 103-127.

Stanley, T.D. and Doucouliagos, H. (2010) "Picture this: a simple graph that reveals much ado about research", *Journal of Economic Surveys* 24(1): 170-191.

Sterne, J.A.C. (2009) *Meta-Analysis in Stata; An Updated Collection from the Stata Journal*. College Station Texas: Stata Press.

Van den Berg, J.C.J.M., Button, K.J., Nijkamp, P. and Pepping, G.C. (1997) *Meta-Analysis in Environmental Economics*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Verbeek, M. (2004) *A Guide to Modern Econometrics*. Chichester West Sussex UK: John Wiley & Sons.

Table 1 Calculation of the Fixed Effects meta-estimate in a simple example

	b	t	se	95% lb	95% ub	v	1/v	w
study A	**-0.1**	**-1.5**	**0.067**	**-0.233**	**0.033**	**0.004**	**225**	**0.707161**
study B	**-0.4**	**-1.7**	**0.235**	**-0.871**	**0.071**	**0.055**	**18**	**0.056769**
study C	**-0.15**	**-1.3**	**0.115**	**-0.381**	**0.081**	**0.013**	**75**	**0.23607**
						sum of v	**sum of 1/v**	**sum of w**
average	**-0.21667**					**0.073**	**318**	**1**
					var FE meta			
FE meta	**-0.129**	**-2.298**	**0.056**	**-0.239**	**-0.019**	**0.003**		

Table 2 Meta-analysis in 10 "easy" steps

Step 1: Collect a large sample of published and unpublished papers that reports statistical estimates on the topic of interest.
Step 2: Choose an appropriate "effect size": a parameter estimate, *t* value, *z* value, correlation coefficient, or an indicator variable (such as "statistically significant at the 5% level").
Step 3: Obtain all effect sizes and the corresponding study characteristics; and code this information in a spreadsheet.
Step 4: Using statistical software, calculate descriptive statistics on the effect size and do plots and cross-tabulations.
Step 5: Carry out the homogeneity test to measure the extent of "excess variation" in effect sizes that must be explained by study characteristics.
Step 6: Select moderator variables that are likely to explain "excess variation" in effect sizes.
Step 7: Carry out a meta-regression analysis (MRA) that regresses effect sizes on moderator variables, taking into account statistical issues associated with MRA.
Step 8: Correct for within-study correlation (clusters) and publication bias, when necessary.
Step 9: Carry out a sensitivity analysis by varying moderator variables and testing statistical properties of the MRA.
Step 10: Write up and publish the results.

Table 3 An example of meta-analysis of categorical findings

Type of fiscal policy	Number of studies	Expected impact	Proportion of studies supporting expected impact	Proportion with inconclusive impact
Education	12	+	0.92	0.08
Infrastructure	39	+	0.72	0.20
Taxation	10	–	0.60	0.40
Defence	21	–	0.52	0.43
Government consumption or "size"	41	–	0.29	0.54
All types	123	As above	0.51	0.36

Source: Nijkamp and Poot (2004)

Figure 1 Confidence intervals for the regression coefficients and the FE meta-estimate in the simple example

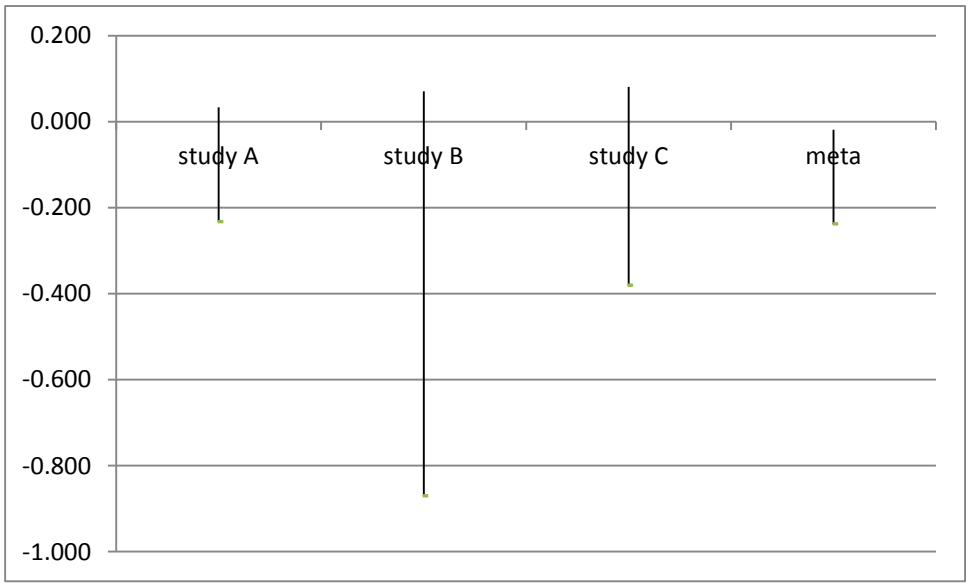

Figure 2 Example of a funnel plot with publication bias