Faculteit der Economische Wetenschappen en Bedrijfskunde

SERIE RESEARCH MEMORANDA

Methodological pitfalls in meta-analysis: Publication bias

Raymond J.C.M. Florax

Research Memorandum 2001-28

July 2001

vrije Universiteit *amsterdam*

METHODOLOGICAL PITFALLS IN META-ANALYSIS: PUBLICATION BIAS*

Raymond J.G.M. Florax

Department of Spatial Economics, Master-point
Free University
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
Phone: +31–20–4446092, Fax: +31–20–4446004
E-mail: rflorax@feweb.nl
URL: http://www.feweb.vu.nl/re/master-point

1. Introduction

An important characteristic of modem science is the enormous productivity of researchers. There is a growing stream of scientific output in the form of patents, publications, and knowledge-based consultancy to industry and the public sector. In this paper we will be concemed with publications, which should in the current context be understood as a rather broad concept. The term 'publication' refers to traditional journal articles (provided as hardcopy or digitally online), monographs, and edited volumes, but also to outlets that are more difficult to access, such as theses and dissertations, research memoranda, working papers, and mimeos of conference papers. Following what is already standard practice in medicine, education, marketing, and psychology, economists now increasingly use meta-analysis as a tool to synthesize and summarize the insights prevailing in the literature (Van den Bergh et al. 1997). The critical feature distinguishing meta-analysis from other types of summarizing techniques, such as state-of-the-art reviews and expert assessments (Button 2001), is its statistical nature. Meta-analysis is concemed with the *statistical* analysis of research results of studies performed previously, and should thus be distinguished from primary and secondary analysis (Glass 1976). Hunter and Schmidt (1990) succinctly define meta-analysis as the 'analysis of analyses'.

Although literature reviews are valuable in their own right, an important drawback is that they are usually implicitly based on vote-counting (Light and Smith 1971). Vote-counting essentially boils down to counting the number of significantly positive, significantly negative, and insignificant results. These results are subsequently simply tallied, and the category with the plurality of cases is usually taken to represent the true characteristics of the underlying population. This procedure is, however, fundamentally flawed because for each estimate there is a probability that the wrong conclusion is drawn (the so-called Type-11 error), and these mistakes do not cancel out when the number of studies considered increases. Consequently, we tend to draw the wrong conclusion more often as the number of studies increases (Hedges and Olkin 1985).

* 1 would like to thank Yasuhiro Sakai and Brigitte Waldorf for helpful comments on an earlier version, and Jasper Dalhuisen and Abay Mulatu for permission to use the databases. Jasper Dalhuisen and Laura Spierdijk have been very helpful in providing research assistance.

Meta-analysis constitutes of a set of techniques that does not necessarily rest on the principle of vote-counting. In meta-analysis, statistical summary indicators of studies performed previously, usually labelled 'effect size', are statistically analysed. Taking into account sign and significance alone — as in the popular vote-counting — is obviously insufficient to determine whether the results of different studies agree. Differences in magnitude of the estimated effects convey important information as well. Moreover, the results of an empirical study may provide a reasonable estimate of the sampling uncertainty of results, but non-sampling issues such as research design, model specification and estimation technique, are usually relatively constant within a study (Hedges 1997). Techniques such as meta-regression, in which non-sampling characteristics can be taken into account as moderator or predictor variables, constitute an attractive and rigorous approach to synthesizing research results.

A paramount methodological problem for meta-analysis is the potentially detrimental effect of publication bias. Publication bias occurs when only studies reporting statistically significant results or with a 'reasonable' magnitude of the effect size are being published, and others are not. This creates a major problem because the selection criterion for publication is a function of the effect size and/or its associated significance level. This phenomenon may be partly the result of self-selection in the behaviour of researchers: research efforts resulting in insignificant results or 'unreliable' effect size estimates are "left in the file drawer" (Rosenthal 1979). The 'publication culture' in which editors of journals only publish significant effect size estimates with the 'right' direction and magnitude of the effect, is likely to be an important determinant of the occurrence of publication bias as well.

Several variations of this problem exist (Greenhouse and Iyengar 1994), although the terminology is not always very clear. One is what Hedges (1990) has labelled 'reporting bias', indicating the tendency to not report statistically insignificant results. The other is 'retrieval bias' (Rosenthal 1990), which — among economists — is more commonly known as 'sample selection bias'. Sample selection bias potentially has a somewhat broader spectrum of underlying causes as compared to reporting or publication bias. If we assume, that the set of published or retrievable studies is a representative sample of the population of studies, selective effects in the sampling of studies for the meta-analysis may stil1 have a negative impact on the validity of the meta-analysis. Selectivity can refer to various aspects of the sampling process: the meta-sample may be biased in terms of, for instance, theoretical perspectives, spatial and/or tempora1 coverage, data type, publication outlet, and statistical techniques. The negative connotation that we usually attach to 'sample selection bias' is indicative of the harmful effect on the validity of the meta-analysis. The latter occurs when there is a systematic relationship between characteristics of the sampling process and the significance or magnitude of the effect size.

The issue of publication bias did not generate a sizeable discussion in the economic literature. Among the few exceptions are Card and Krueger (1995), and Ashenfelter et al. (1999), who systema-

are often considerable: for instance, in residential water demand studies, price elasticities reported in the literature range from -7.5 to +7.9, and income elasticities vary between -0.9 and +7.8 (Dalhuisen et al. 2001). Prospective important factors causing this variation include differing theoretical and modelling perspectives, and differences in research designs (such as, time-series or panel data, survey or non-survey information), but also behavioural aspects, such as population density, geographical location, and income differentials.

It should, however, be pointed out that several methodological pitfalls may invalidate the conclusions of a meta-analysis, or at least evoke considerable scepticism. Glass et al. (1981) distinguish four types of methodological problems:

- empirical results, which turn out not to be significant in a statistical sense, are rarely published;
- overall conclusions may not be warranted, due to the comparison and aggregation of studies that employ different measuring techniques, different variables, and the like;
- poorly designed studies are not treated differently from well-designed studies; and
- multiple results from the same study are often used, possibly biasing or invalidating the meta-analysis due to lacking independence of observations.

These methodological problems can be rephrased into three methodological requirements for a proper meta-analysis: the sample selection and the publication process should be free of bias, the effect sizes observed in the meta-sample should be homogeneous, and the observed effect sizes in the meta-sample should be independent. The latter is especially doubtful in the case of multiple sampling of effect sizes from the same study. If the above conditions are not met, appropriate solutions or correction mechanism should be employed. The extent to which the abovementioned methodological pitfalls have been adequately treated, both in terms of detection and remediation, varies widely in environmental economic meta-analyses.

It is easy to see that the homogeneity requirement is usually violated — in largely non-experimental sciences, such as economics, probably to an even greater extent than in the more uniform experimental sciences. The heterogeneity may show up in two different ways. One is in the form of differences in research design and spatio-tempora1 characteristics of the primary analyses. These are usually adequately treated as fixed effect differences in a regression framework.[2] Heterogeneity may also show up as heteroscedasticity, which is intrinsic to meta-analysis because the underlying studies usually have different sample sizes, and hence sampling variance.

In (environmental) economics, meta-analyses are almost invariably based on multiple sampling from the same study, among other things because replication is not very popular in

[2] Alternatively, these differences can be modelled as random effects, but that is the exception rather than the rule in environmental economics. One should note, however, that the extent to which the results of a meta-analysis can be generalized is crucially different between fixed and random effects models (see, e.g., Hedges and Olkin 1985).

tically investigate the occurrence and impact of publication bias with respect to studies on minimum wages, and studies on the relation between schooling and eamings, respectively. In the area of environmental economics, specifically in the field of environmental valuation that constitutes the prime area in which meta-analysis has been applied, publication bias received some attention as well.[1] Smith and Huang (1995), for instance, stress the disturbing effect that sample selection bias may have on the outcome of the meta-analysis. They use a two-stage Heckman-like probit model to determine the likelihood of sample selection bias, and subsequently include the inverse Mill's ratio in the meta-regression. The ratio is related to the estimated probability of including a study in the meta-sample on the basis of the year to which the data refer, the use of actual prices, and the significance and direction of the coefficient for pollution.

This paper is concemed with publication bias as an important methodological pitfall in meta-analysis. We will discuss conceptual issues related to publication bias and sample selection, describe techniques to identify and remedy publication bias, and provide some illustrations of these techniques. The organization of the remainder of this paper is as follows. Section 2 positions the issues of publication bias and sample selection in the broader context of methodological pitfalls of meta-analysis. We conclude that both lacking independence of effect sizes sampled from the same study, and publication bias are practically ignored in meta-analyses in economics. In Section 3, various techniques to detect publication and sample selection bias are introduced. These techniques range from eyeball assessment of so-called funnel graphs to rather complex econometric models. Section 4 gives an overview of the use of most of these techniques in the context of environmental economics. Three meta-databases, dealing with price and income variability of residential water demand, and the impact of stringency of environmental policy on intemational trade flows, are used for illustrative purposes. Section 5 winds up this paper and summarizes the main conclusions.

2. Methodological pitfalls in meta-analysis

Meta-analysis has an incredible appeal as well as a promising potential for applied studies in the field of environmental economics. In particular in studies focussing on the economic valuation of environmental degradation or improvement, it is useful to investigate whether a common effect size exists that can be used by policymakers when deciding on policy options for unstudied policy sites (in a so-called value or benefit transfer). In addition, meta-analysis can be used to explore the factors that are influential in explaining variations in point estimators among individual studies. Such variations

[1] Meta-analyses on urban pollution valuation studies are carried out by, for instance, Schwartz (1994), Smith (1989), Smith and Huang (1993, 1995), and Van den Bergh et al. (1997). on recreational benefits by, e.g., Smith and Kaoru (1990), Smith and Osborne (1996). Sturtevant et al. (1995). and Walsh et al. (1989). and on groundwater and wetland valuation by, e.g., Boyle et al. (1994). Brouwer et al. (1997). and Woodward and Wui (2001).

economics. As a consequence, effect sizes cannot be considered independent. In an observational sense, it may be difficult to distinguish heterogeneity from dependence: a clustering of similar values sampled from the same study can either be viewed as heterogeneity or as dependence among effect sizes sampled from the same primary study. There is to date no meta-analysis in (environmental) economics that treats independence as a relevant problem. The more sophisticated studies invariably focus on heterogeneity, which is usually taken into account by means of fixed effects, sometimes in combination with a heteroscedasticity-robust estimator (see, e.g., Smith and Osborne 1996).

The problem of sample selection and/or publication bias is for the most part practically ignored. There are a few meta-analyses in which a fixed effect is included to distinguish between different publication outlets. Usually 'published' monographs, edited volumes and journal articles are contrasted with 'unpublished' theses and dissertations, research memoranda, working papers, and mimeos of conference papers.[3] This constitutes, however, a rather crude representation of the publication selection process. The coefficient of the fixed effect will merely signal whether a *p*-value-related or a size-effect-related selection process (in an ordered probit set-up or a continuous regression set-up, respectively) is apparent in the meta-sample. The selection of studies from the wider set of retrievable studies is implicitly stil1 assumed to be free of bias.

There are two important additional limitations to this approach. First, the published-unpublished distinction may be rather artificial. because the categorization is time-dependent: a working paper may at some later point in time be published in a journal or edited volume. Second, the definition of what is considered 'published' is highly arbifrary: compare, for instance, an 'unpublished' but refereed working paper at a top-notch university to an article 'published' in a weakly refereed, rather obscure, journal.

Smith and Huang (1995) are a noteworthy exception, in the sense that they go beyond the typical published-unpublished distinction and consider the sample selection process as well. On the one hand, they operationalize the above distinction between 'published' and 'unpublished' studies by means of a fixed effect. On the other, however, they explicitly investigate the sampling process underlying their meta-analysis sample by means of a two-stage Heckman procedure. Both the (narrowly defined) publication bias as well as the sample selection bias is present in their analysis of hedonic estimates of air quality. An important drawback of this approach is of course that inclusion of all retrieved studies in the meta-sample is precluded, because the modelling of the sample selection process is based on distinguishing studies included in the meta-sample from those that are not included. The latter may be based on rather arbitrary criteria, and in a sense it also shows that the sample selection problem is 'shifted' rather than fully taken into account. In the two-stage Heckman approach, one stil1 has to assume that the sample of studies used for the analysis, comprising both the

[3] For instance, in Van den Bergh et al. (1997, pp. 130-132) it is shown that multiplier effects in tourist regions are generally lower for estimates published in scientific journals.

studies to be included in the meta-sample as well as the studies used in the selection stage only, is representative of the population and/or the full set of retrievable studies.

3. Detecting and remedying publication bias

Although methods to detect and remedy publication bias are not yet widely used in environmental economic meta-analyses, a substantial arsenal of methods is available. Methods to detect and remedy publication bias range from the mere avoidance of sample selection bias, and quasi-statistical techniques, to more rigorous statistical methods. Publication bias is essentially a result of selective sampling. The selection effect can be the consequence of a publication process that is biased towards either the magnitude of the effect size or the p-value, or both. The methods, concisely summarized below, tend to concentrate on one or the other possible cause for publication bias.[4] In addition, many of the available techniques only focus on the *detection* of publication bias, leaving the researcher in the blind as to the exact magnitude of the bias and the impact on the analysis of effect sizes.

Below, we will discuss several of the techniques, which can be grouped into three genera1 classes. The first class is in fact concemed with the avoidance of publication bias through the use of appropriate sampling frames. The second class of techniques centres on the detection of publication bias, and comprises several univariate and bivariate test statistics. Finally, the third class of techniques has a (multivariate) regression framework in common. These techniques take into account the publication and/or sample selection process, and the results of the meta-analysis are hence — to varying degrees — robust to publication bias.

3.1 *Sampling frames*

An obvious, although fairly tedious, approach is to retrieve all studies, published as well as unpublished. This approach is appealing and, in a theoretical sense, the most favourable one to address, and in effect even estimate, publication bias. It is, however, severely hampered by the several problems. There is of course no way to ensure that all unpublished results (e.g., in languages foreign to the investigator) are taken into account. Moreover, unpublished studies making up the so-called 'fugitive literature' (Rosenthal 1994), which is oftentimes poorly documented and referenced, are usually difficult to acquire. In the sciences, in particular in medicine, these problems are — at least partially — remedied through the development of registries of clinical trials. Registries will increasingly facilitate literature retrieval, and may thus be expected to lead to an increase in the number and scope of meta-analyses in this field of research (Petitti 1994).[5]

---

[4] Begg (1994) maintains that a methodology accounting for both types of effects is not yet available. but we already saw that the two-stage selection model of Heckman (1979) can be used to account for both.

[5] Registers of non-experimental studies, which are often based on secondary analysis of data collected for other purposes, have not yet been created (Petitti 1994).

3.2 A quasi-statistical graphical technique

If the gathering of all studies is not feasible or not efficient, one can turn to statistical or quasi-statistical techniques to detect publication bias. A quasi-statistical technique, introduced by Light and Pillemer (1984), is a graphical analysis where the effect size estimates are plotted on the horizontal and the sample size of the respective studies on the vertical axis. Distortions of a funnel like shape (with the tip pointed up, and centred around the 'true' effect size under the null hypothesis of no publication bias) may be taken as an indication that publication bias is present. The distortions can of course be several, and it is not always clear what causes the distortions. The well-known selection effect on the basis of significance and size is signalled by a graph that is skewed to the right or left, or with the lower centre part missing.[6]

Obviously, this method is not very precise, as there is a good deal of subjective judgment required in determining distortions of the funnel-like shape. Although it is based on the statistical property that the variance of the effect size is roughly inversely proportional to sample size, inferences from a graphical analysis do not really have a rigorous statistical basis. Furthermore, other (unknown) factors may be responsible for distortions from the hypothesized funnel-like appearance. If the meta-analysis sample contains relatively few studies this approach may, however, be the only feasible altemative.

In econornics, many crucial statistical summary indicators that can be used as effect sizes, such as elasticities and multipliers, are defined to be strictly positive or negative (eventually including zero). For instance, price elasticities of demand for a normal good are defined to be negative. Positive elasticity estimates are therefore rare, which distorts the funnel-like shape through right-censoring. Because a positive relationship between the price of water and demand is extremely rare in practice, this distortion of a funnel-like shape is not necessarily indicative of publication bias.

3.3 *The file drawer test*

The test developed by Rosenthal (1979) is generally referred to as the 'file drawer test'. The intuitive idea behind it is simply to calculate the number of 'left-in-the-file-drawer' studies with non-significant

---

[6] For a difference in means between an experimental (M^E) and a control (M^C) group, standardized by some standard deviation s, we know that:

$$(M^E - M^C)/(s/\sqrt{n}) = \theta/\sqrt{n}$$

follows a t-distribution, from which we can infer the relationship between the effect size θ and the probability value: for a given effect size magnitude the higher n the lower the p-value, and for a given n the higher the effect size the lower the p-value. The same holds for effect sizes defined as elasticities. From a doublelog specification of a regression model we can take the elasticity value b and the estimated standard error s_b, and b/s_b follows a t-distribution. The same funnel-like shape should be apparent under the null hypothesis of no publication bias because the elasticity value is equal to the effect size ($b \equiv \theta$), and the standard error is roughly inverse proportional to the square root of sample size.

p-values, on the basis of a combined test on the k studies with significant results and the k_0 unpublished studies. The combined significance test is:

$$z = \sum_{i=1}^{k} \frac{\Phi^{-1}(p_i)}{\sqrt{k + k_0}} = \frac{\sqrt{k}\, z_0}{\sqrt{k + k_0}} \tag{1}$$

where p_i is the one-tailed p-value for the ith study, and $\Phi(\ )$ the cumulative standard normal distribution. Substituting z with some desired critical value of the normal distribution C_α, and subsequent re-arranging, leads to an estimate of k_0:

$$k_0 = k(z_0^2 - C_\alpha^2)/C_\alpha^2 \tag{2}$$

Whenever the number of unpublished studies with (assumingly) null results is large enough, the researcher may be confident that the outcome of the meta-analysis is not due to selective sampling of studies with significant results. A smal1 number of k_0 implies that a fairly smal1 number of unpublished studies could overthrow the conclusion based on the meta-analysis of the published studies.

An obvious drawback of the file drawer approach is the use of a test that combines study results by means of probability values (Hedges and Olkin 1985, p. 306). The altemative hypothesis of such a test is not necessarily very informative, because rejection of the null hypothesis that the combined effect size for all studies is unequal to zero merely implies that there is at least *one* study that has a nonzero effect. This drawback is epitomized by the fact that the reasoning on which the file drawer test rests, relies on the assumption that the results of the unpublished studies are in effect equal to zero (Hunter and Schmidt 1990, p. 512).[7] Orwin (1983) presents a slightly less strict formulation of the test, and uses the criterion of selection on the basis of the magnitude of the effect size. He looks for the number of null studies needed to reduce the average effect size estimate to a negligible quantity.[8]

3.4 Concordance tests of effect size

A statistical procedure that does not rely on the questionable modelling assumption of zero (unpublished) effect sizes can be based on a pairwise rank-ordering of two factors, such as effect size and sampling variance, so that a test on publication bias may be obtained by using Kendall's τ or Spearman's ρ. A main disadvantage of this type of tests is, however, their lack of power (Begg 1994).

---

[7] Petitti (1994, pp. 129-130) reviews some specific drawbacks for medical studies.

[8] The history of this variant goes in fact back to 1979, when Hunter and Schmidt (1990, p. 512) originally developed it.

The test is based on ranking the standardized effect sizes $\{ T_i^* \}$, assuming that they will be independently and identically normally distributed, versus the sampling variances $\{ v_i \}$ or the sample sizes $\{ n_i \}$. The effect size of study i can be standardized as follows:

$$T_i^{\bullet} = \frac{T_i - \overline{T}_{\bullet}}{\sqrt{\tilde{v}_i}} \tag{3}$$

with

$$\overline{T}_{\bullet} = \frac{\sum_{i=1}^{k} v_i^{-1} T_i}{\sum_{i=1}^{k} v_i^{-1}}$$

and

$$\tilde{v}_i = v_i - \left(\sum_{i=1}^{k} v_i^{-1} \right)^{-1},$$

where the latter represents the variance of $(T_i - \overline{T}_{\bullet})$. A normalized z-value can then be obtained, involving P, the number of all possible pairings in which one factor is ranked in the same order as the other, and Q, the number in which the ordering is reversed, by means of:

$$z = \frac{P - Q}{\sqrt{k(k-1) \cdot (2k + 5)/18}} \tag{4}$$

where k is the total number of studies in the meta-analysis. Begg (1994, p. 403) suggest a plot of $T_i^{\bullet}$ versus $\sqrt{v_i}$ or n_i in order to graphically determine publication bias, as one has to detect mere correlation instead of a funnel effect in such a graph.[9]

3.5 Weighted distribution theory and selection on the basis of p-values

The statistically most rigorous approach, which is stil1 in the process of being developed, is based on the assumption that each study i, with an estimated statistic X_i, can be assigned a weight function

[9] Begg (1994, p. 107) also suggests a rank correlation test, as described in the current subsection, involving the estimated weights and the estimated probabilities of the step function, using the weighted distribution approach described in the next subsection.

$w(X_i)$, which determines the probability of being observed, i.e., of being published. Until now, various authors have used the assumption that the weight function is determined by the p-value, rather than the effect size (Begg 1994, p. 406).[10] Hedges (1992, p. 249) presents a detailed justification for employing the probability value as the determinant of the weight function, making reference to psychological research on the interpretation of statistical analyses.

The publication selection process is modelled using weights by means of a step function, with *a* priori determined discontinuities. Several variants have been suggested in the literature. In Lane and Dunlap (1978) and Hedges (1984) discontinuities are introduced by assigning a weight of 1 to significant results and 0 to others. Hedges (1992) uses weights according to the scheme $p < 0.01$, 0.01 c $p < 0.05$, and $p > 0.05$. Iyengar and Greenhouse (1988) specify a weight function in which the probability of being observed is 1 for $p < 0.05$, and the remaining weights decline exponentially or are constant (though not 1). Finally, Dear and Begg (1992) estimate the discontinuities from the data. Below, the method suggested by Hedges (1992) is followed, but there is no loss of generality. Hedges' (1992) analysis is based on p-values of a two-sided test of the effect size being different from zero. The variant using one-sided p-values is described in Vevea and Hedges (1995). Although in economics a one-tailed pattem of selection is usually more appropriate, Vevea and Hedges (1995, p. 424) observe that as the population effect grows larger (in absolute value), the contribution of the negative (or positive) tail of the distribution becomes negligible. As a consequence, one-tailed and two-tailed selection models oftentimes yield essentially equivalent results."

The following notation is introduced: let $\{X_i\}$ be a set of effect size variables from i different studies such that $X_i \sim N(\delta_i, \sigma_i^2)$, where δ_i is an unknown, normally distributed parameter with unknown mean A, and variance σ^2. Hence, it follows that the observed effect size X_i follows a normal distribution with an unknown mean A, and variance $(\sigma_i^2 + \sigma^2)$, where the unknown mean Δ_i can be modelled as a function of linear predictors (for instance, $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$). The observed test statistic X_i from the primary study i tests the null hypothesis that $\delta_i = 0$ by means of the test statistics $Z_i = |X_i| / \sigma_i$, which is associated with the two-tailed p-value $1 - \Phi(Z_i) + \Phi(-Z_i)$

Hedges (1992) introduces the following weighting scheme, where the weights ω_i represent relative probabilities because one of the weights is fixed to an arbitrary value:

---

[10] In terms of the effect size, each study has a different weight function, unless all studies have the same sample size, and hence, conditional variance Begg (1994, p. 106).

[11] A one-tailed selection model is easily derived from the two-tailed model given here, and is presented in Vevea and Hedges (1995).

$$w(p_i)=\begin{cases}\omega_i & \text{if } 0 < p_i \le a_1,\\ \omega_j & \text{if } a_{j-1} < p_i \le a_j,\\ \omega_k & \text{if } a_{k-1} < p_i \le 1.\end{cases} \tag{5}$$

where a refers to the *a priori* determined endpoints. A logica1 choice is to set ω_i = 1.0 . This constraint implies that the ω_i -values represent the chance that an estimate with a given p-value is observed relative to the chance that studies with $p \le a_1$ are observed. Because the p-values depend on both X_i and σ_i^2 , and they are assumed to follow a norrnal distribution, the weight function as a function of X_i reads as:[12]

$$w(X_i,\sigma_i^2)=\begin{cases}\omega_1 & \text{if } -\sigma_i\Phi^{-1}(a_1/2) < X_i \le \infty \text{ and } X_i > 0,\\ \omega_j & \text{if } -\sigma_i\Phi^{-1}(a_j/2) < X_i \le -\sigma_i\Phi^{-1}(a_{j-1}/2) \text{and } X_i > 0,\\ \omega_k & \text{if } 0 < X_i \le -\sigma_i\Phi^{-1}(a_{k-1}/2),\\ \omega_1 & \text{if } -\infty \le X_i < \sigma_i\Phi^{-1}(a_1/2) \text{and } X_i < 0,\\ \omega_j & \text{if } \sigma_i\Phi^{-1}(a_{j-1}/2) \le X_i < \sigma_i\Phi^{-1}(a_j/2) \text{and } X_i < 0,\\ \omega_k & \text{if } \sigma_i\Phi^{-1}(a_{k-1}/2) \le X_i < 0.\end{cases} \tag{6}$$

where $\Phi^{-1}(p)$ is the inverse norrnal cumulative distribution function evaluated at p.

The weighted probability density of X_i given the weight function $w(X_i, \sigma_i^2)$ and the parameters σ_i^2, $\beta=(\beta_0,\beta_1,\ldots,\beta_p)'$ and $\omega=(\omega_1,\ldots,\omega_k)'$ is:

$$f(X_i|\beta,\sigma^2,\omega)=\frac{w(X_i,\sigma_i^2)\phi\left(\dfrac{X_i-\Delta_i}{\eta_i}\right)}{\eta_i A_i(\Delta_i,\eta_i^2,\omega)}, \tag{7}$$

where A_i is the sum of norrnal integrals over the regions where the weight function is constant, which may be expressed as:

$$A_i(\Delta_i,\eta_i^2,\omega)=\int_{-\infty}^{\infty}\eta_i^{-1}w(X_i,\sigma_i^2)\phi\left(\frac{X_i-\Delta_i}{\eta_i}\right)\mathrm{d}X_i, \tag{8}$$

[12] Note that the weight function described here assumes a two-sided test. The one- or two-sidedness depends on the validity in the context of publication sampling, and does not necessarily refer to the characteristics of the test in the original studies.

where $\eta_i^2 = \sigma_i^2 + \sigma^2$, ϕ is the standard normal density function, and $\Delta_i = X_i\beta$.

On the basis of the individual likelihoods for the independent observed data $X = (X_1, \ldots, X_n)'$ of the original studies, the joint likelihood is:

$$\ell(\beta, \sigma^2, \omega | X) = \prod_{i=1}^{n} \frac{w(X_i, \sigma_i^2)\phi\left(\frac{X_i - \Delta_i}{\eta_i}\right)}{\eta_i A_i(\Delta_i, \eta_i^2, \varpi)} \tag{9}$$

Hedges (1992) derives the log-likelihood:

$$L = c + \sum_{i=1}^{n} \log w_i(X_i, \sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \Delta_i}{\eta_i}\right)^2 - \sum_{i=1}^{n}\log(\eta_i) - \sum_{i=1}^{n}\log\left(\sum_{j=1}^{k}\omega_j B_{ij}(\Delta_i, \sigma^2)\right), \tag{10}$$

where $B_{ij}(A, , \sigma^2)$ is the probability that a normally distributed random variable, with mean Δ_i and variance η_i^2, is assigned a specific weight value. That is:

$$B_{ij} = \begin{cases} 1 - \Phi\left(\frac{b_{i1} - \Delta_i}{\eta_i}\right) + \Phi\left(\frac{-b_{i1} - \Delta_i}{\eta_i}\right) & \text{if } j = 1, \\ \Phi\left(\frac{b_{ij-1} - \Delta_i}{\eta_i}\right) - \Phi\left(\frac{b_{ij} - \Delta_i}{\eta_i}\right) + \Phi\left(\frac{-b_{ij} - \Delta_i}{\eta_i}\right) - \Phi\left(\frac{-b_{ij-1} - \Delta_i}{\eta_i}\right) & \text{if } 1 < j < k, \\ \Phi\left(\frac{b_{ik-1} - \Delta_i}{\eta_i}\right) - \Phi\left(\frac{-b_{ik-1} - \Delta_i}{\eta_i}\right) & \text{if } j = k, \end{cases} \tag{11}$$

where b_{ij} denotes the left endpoints of the intervals of positive X values assigned weight ω_j in the ith study, that is, $b_{ij} = -\sigma_i \Phi^{-1}(a_j / 2)$.

Hedges (1992) presents the first and second derivatives for this log-likelihood, and gives suggestions for the computational procedures to be followed in estimation. In addition, two tests to detect possible publication bias are suggested.

The first test is a χ^2 Pearson test based on grouped frequencies. The test has k-1 degrees of freedom, and reveals the goodness of fit of the observed p-values to the expected p-value distribution. Assume j intervals defined by the cut-off points $0 \equiv a_0 < a_1 < \ldots < a_k \equiv 1$, and count the observed number O_j of p-values in the jth interval $[a_{j-1}, a_j]$, and estimate the expected number E_j of p-values in the same interval, using:

$$E_j = n \int_{a_{j-1}}^{a_j} f\left(p|\sigma_1,\ldots,\sigma_n,\hat{\Delta}_0,\hat{\sigma}_0\right)dp = \sum_{i=1}^{n} B_{ij}\left(\hat{\Delta}_0,\hat{\sigma}_0\right), \tag{12}$$

where B_{ij} is as given above, and the subscript 0 refers to the situation in which there is no publication bias, and hence ϖ_2 = . . . = ϖ_k = 1. The Pearson goodness of fit test statistic is given by:

$$\sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2(k-1) \tag{13}$$

for the null hypothesis that there is no publication bias.

The second test is a straightforward Likelihood Ratio (LR) test, which takes the usual form, that is:

$$2\left(L\left(\hat{\Delta},\hat{\sigma},\hat{\varpi}\right) - L\left(\hat{\Delta}_0,\hat{\sigma}_0,\hat{\varpi}_0\right)\right) \sim \chi^2(k-1) \tag{14}$$

and compares the unrestricted and the restricted maximum likelihood estimates for differences in fit among different specifications with different constrained parameters. A test on publication bias results if in the constrained model the vector of estimated weights is restricted to be a ***k-1*** unity vector.

3.6 A two-stage Heckman approach

A detailed treatment of the two-stage Heckman approach to sample selection (or 'incidental truncation') is beyond the scope of this paper. The literature on this subject is very extensive (see, e.g., Heckman 1990, for a review). The basic idea is, however, rather straightforward, because the bias resulting from the use of non-randomly selected samples is comparable to the ordinary problem of omitted variables (Heckman 1979, p. 155). This can be seen as follows (see, e.g., Greene 1993, pp. 708-7 10). Assume that we are interested in the meta-equation:

$$y_i = \beta' x_i + \varepsilon_i \tag{15}$$

where y_i is the effect size measure, and x, a vector containing variables explaining the variation in the observed effect sizes. The effect size is, however, only observed if the selection variable z_i = 1. The selection mechanism is modelled as:

$$
\begin{aligned}
z_i^* &= \gamma' s_i + u_i \\
z_i &= 1 \quad \text{if } z_i^* > 0 \\
z_i &= 0 \quad \text{if } z_i^* \le 0 \\
\Pr(z_i = 1) &= \Phi(\gamma' s) \\
\Pr(z_i = 0) &= 1 - \Phi(\gamma' s)
\end{aligned}
\tag{16}
$$

where s_i is a vector of variables influencing the selection. The conditional mean for the observed effect sizes is then:

$$
\begin{aligned}
E[y_i \mid z_i = 1] &= E[y_i \mid z_i^* > 0] \\
&= \beta' x_i + \beta_\lambda \lambda_i(\alpha_u)
\end{aligned}
\tag{17}
$$

where

$$
\alpha_u = \frac{-\gamma' s_i}{\sigma_u} \quad \text{and} \quad \lambda(\alpha) = \frac{\phi(\gamma' s_i / \sigma_u)}{\Phi(\gamma' s_i / \sigma_u)}
$$

with $\lambda(\alpha)$ being referred to as the inverse of Mill's ratio. Given equation (17) it is obvious that estimating equation (15) produces inconsistent estimates of β, due to an omitted variable problem. Consistent estimates for the meta-model can only be obtained when both x and λ are included as regressors.

An obvious advantage of this sample selection approach is that it allows a detailed analysis of the sample (or publication) selection process. This approach goes beyond the evidently simpler approach based on weighted distribution theory, which merely considers selection on the basis of *p*-values. A disadvantage of the sample selection approach is that not all retrieved studies can be included in the meta-analysis.

4. Illustrations in environmental economics

We will demonstrate the use of the abovementioned techniques, except for the two-stage selection approach, for two examples from environmental economics. One is concemed with a study of price and income elasticities of residential water demand, and the other deals with the impact of strictness of environmental policy on intemational trade flows. The data for these examples are taken from two recent meta-analyses in environmental and natural resource economics, extensively documented in

Dalhuisen et al. (2001) and Mulatu et al. (2001).[13] Instead of employing the full samples used in these studies, we restrict the selection of the meta-sample observations to effect sizes defined as elasticities. In addition, we require probability information on a test of the elasticity being significantly different from zero to be available. Taking into account these two restrictions, the following meta-samples are available for illustrative purposes:

- a sample of 110 price elasticities (mean -0.38, standard deviation 0.41), derived from 24 studies, with 77 elasticities being significantly different from zero based on a one-sided test of the elasticity being negative at the 0.01 level;
- a sample of 90 income elasticities (mean 0.35, standard deviation 0.45), derived from 17 studies, with 48 elasticities being significantly different from zero based on a one-sided test of the elasticity being positive at the 0.01 level;
- a sample of 103 stringency elasticities (mean -0.46, standard deviation 2.31), taken from 4 studies, with 34 elasticities being significantly different from zero based on a one-sided test of the elasticity being negative at the 0.01 level.

We do not intend to give an overview of the pivotal issues in the literature on residential water demand and environmental regulation and competitiveness, respectively, nor do we use very elaborate and adequate specifications with fixed effects accounting for differences among studies. The examples are therefore merely illustrations of how the publication bias techniques can be fruitfully applied, and no substantive conclusions will be drawn regarding the issues at stake in the literature.

One important proviso should be made at the outset of the analysis. Most of the techniques to assess (and correct for) publication bias are based on the assumption of independent and identically distributed effect sizes. In economics, as in many other non-experimental sciences, the number of available studies is rather limited, and most studies report empirical estimates for various different specifications. In order to obtain a sufficient number of observations for a meta-analysis, multiple sampling per study is the rule rather than the exception. Obviously, the estimated effect sizes are then dependent, among other things because they have been estimated using the same data. This problem has not been extensively treated in the methodological meta-analytical literature, and is therefore disregarded in the examples reported below.

Figure 1 presents the price, income and stringency elasticities, ordered according to magnitude and plotted in deciles of the available meta-samples. It is obvious from Figure 1 that the price elasticities are largely negative, and the income elasticities positive, with a relatively smal1 standard deviation. For stringency elasticities the division in positive and negative elasticities is much more

[13] The papers and complete databases are available online at http://www.tinbergen.nl (see 'Publications') and http://www.econ.vu.nYre/master-point (see 'Download'). respectively.

even, and the standard deviation is considerably larger (mainly due to a few large (in absolute value) negative observations).

A preliminary indication of publication bias can be taken from the funnel graphs. Figure 2 presents such graphs for the different elasticities, with control lines added that should roughly contain 90% of the estimates (following Vevea and Hedges 1995). For price elasticities it is evident that there is right-censoring, which is plausible given that water is a normal good, so negative price elasticities are to be expected. However, it also seems as if larger effect sizes (in absolute value) with relatively low p-values are overrepresented, because the graph is slightly skewed to the left. Regarding income elasticities the former phenomenon can be observed as well. The censoring is on the left hand side in this case, which is in accordance with the a priori theoretical expectation that water is a normal rather than a Giffen good. For stringency elasticities the theoretical expectation of the direction of the effect is negative if we follow neoclassical theory, but the Porter hypothesis suggests that positive effects of stringent environmental policy on intemational trade flows can be expected (see Mulatu et al. 2001, for details). The graph shows a funnel shape, although it is again not perfectly centred around the sample mean and there is a selection effect with regard to (in absolute value) large negative effect sizes.[14] In sum, it seems that in all cases relatively large effect sizes (in absolute value) are over-represented, which provides evidence for a one-tailed selection process. For price (and stringency) elasticities positive values are censored, and for income elasticities negative values. In that sense, there is evidence for publication bias.

The funnel graph technique essentially assists in detecting biased sampling on the basis of the magnitude of the effect size estimates. Given the statistical property that the estimated standard error of the effect size is roughly inversely proportional to sample size, the funnel graph also provides some insight into biased sampling on the basis of p-values. Significant effect size estimates are likely to be clustered towards the top of the funnel and alongside the edges.

A helpful graph that avoids using this rather cumbersome interpretation is presented in Figure 3, where the standardized effect size is plotted against the estimated standard error (following up on Begg's (1994, p. 403) suggestion). Because of the standardization of the effect size by means of the estimated standard error, no judgmental evidence of a funnel-like shape is necessary, and one can resort to checking the correlation. Under the null hypothesis of no publication bias, the graph should not have a funnel-like shape, but the points should instead appear as if they were randomly allocated over the surface. Figure 3 shows that this is evidently not the case, for neither of the elasticities, and thus selection on the basis of magnitude of the effect size seems plausible.

The file drawer test shows considerable evidence that publication bias is *not* present. The test

[14] An elaborate explanation of the way in which estimated standard errors have been calculated is given in Mulatu et al. (2001).

is based on the combination of one-sided p-values.[15] The results for the combined (one-sided) z-tests are: -32.91 for price elasticities, 23.64 for income elasticities, and -10.83 for stringency elasticities, which are all highly significant (p c 0.01). The null hypothesis of the elasticity being zero is thus rejected on the basis of the combined information of the different studies, but (as mentioned earlier) statistically this merely implies that at least one study has a non-zero effect. The number of unpublished studies with null results that one would need to overthrow this conclusion is of course correspondingly high: 43,909 studies for price elasticities, 18,492 studies for income elasticities, and 4,359 studies for stringency elasticities. The file drawer test therefore leads to the inference that it is highly unlikely that publication bias exists.

The concordance test leads to a slightly different conclusion. The obtained z-values are 0.97 (p = 0.33), -1.46 (p = 0.07), and -1.43 (p = 0.08) for price, income, and stringency elasticities, respectively. This implies that for income and stringency elasticities significant correlation between the pairings occurs, and publication bias is therefore likely to be present.[16]

As a final illustration, we provide results based on the weighted distribution approach of Vevea and Hedges (1995), using an executable binary provided by the authors.[17] A first result is the Pearson χ^2 test, with k-1 degrees of freedom, comparing the observed and expected number of p-values in (k) exogenously determined discrete intervals, under the null hypothesis of no publication bias. The test is given in equation (13), and the numerical results are presented in Table 1. One should note that the tests on the effect sizes are one-sided tests on the effect size being negative for price and stringency elasticities, and the effect size being positive for income elasticities.

The Pearson test in Table 1 shows that publication bias is present in all three meta-samples, and it is lowest for the income elasticities. Positive price elasticities, negative income elasticities, and positive stringency elasticities (p > 0.50) are evidently underrepresented in the respective samples. For both price and stringency elasticities the underrepresentation is, however, not limited to effect sizes with the 'wrong' sign, but it extends to a larger group of effect sizes with relatively large p-values. In addition, highly significant effect sizes are clearly overrepresented for both price and stringency elasticities.

Subsequently, we estimate sample selection models according to the Vevea and Hedges (1995) framework. The results for income and stringency elasticities are presented in Tables 2 and 3. No results are available for the price elasticity meta-sample as the maximum likelihood routine fails to converge, which is likely to be caused by the disproportionate number of very small p-values in this meta-sample. Tables 2A and 3A show the estimation results of the random effects estimator, with and

[15] In order to avoid excessively smal1 and large p-values, minimum and maximum bounds on the p-values were set at 0.000001 and 0.999999 for the file drawer test.

[16] For the concordance test, ties in ranking are crucial. Ties are determined with a precision of six digits.

[17] Jack Vevea (jvevea@email.unc.edu) kindly provided the software.

without additional predictor variables, and with and without the correction for selective sampling. Tables 2B and 3B show the estimated mean elasticity values for different categories, which can be deducted from Tables 2A and 3A and additional information on the covariance between the estimated parameters provided in the variante-covariance matrix (which is not given here, but available on request).

Table 2A clearly shows a pattem consistent with publication selection on the basis of *p*-values. Positive and significant effect sizes are more likely to be included: the estimated weights are almost monotonically decreasing with increasing p-values, as can be seen in Figure 4. The LR test for selection effects is also highly significant. In a fixed effects model (not presented here) the common effect is 0.35 with an estimated standard error of 0.05. This is clearly different from the common effect in a random effects setting, which is estimated to be 0.27, with a highly significant between-studies variance component estimate. However, when selection effects are taken into account, it tums out that the common effect reduces to approximately zero.

The last two columns of Table 2A provide evidence for differences in elasticities on the basis of different underlying models used in the primary studies. The omitted category represents those studies that use average or fixed prices to estimate the demand function. Some studies, however, use marginal or Shin prices, and the studies also differ with respect to the inclusion of a so-called difference variable and the use of a discrete/continuous choice approach (see Dalhuisen et al. 2001 for details). In order to facilitate interpretation, Table 2B presents the conditional means and standard errors for the different types of elasticities. It shows, that when selection effects are accounted for, elasticities based on average and marginal prices are no longer significantly different from zero, unless they are based on a discrete-continuous choice approach. Elasticities based on the latter approach as well as those based on the use of Shin prices are significantly different from zero.

Table 3A and Figure 4 show the results for the meta-analysis of stringency elasticities. The results for the LR test provide evidence for selection effects, but overall the results are rather awkward, as they seem to indicate that the probability of including studies with insignificant p-values is more likely than the inclusion of studies in the first (most significant) interval, especially for negative stringency elasticities ($p < 0.50$).[18] This may be partly due to lacking robustness of the selection model. Simulation experiments have shown that the selection model's ability to reduce the bias of effect size estimates when censorship has occurred is not very robust to violations of the assumed normal distribution for random effects when the between-studies variance component is large compared with the conditional variance (Vevea and Hedges 1995. p. 432). The latter is the case for the stringency elasticity meta-sample.

[18] Vevea and Hedges (1995, p. 430) note that this can occur only if there are fewer studies than expected in the first interval, which is the case here (see Table 1).

In a fixed effects setting (not presented here) the estimated common effect is -0.46, with an estimated standard error of 0.23. The common effect estimate in a random effects setting corrected for selection bias is considerably greater in absolute value (-1.54, see Table 3A). In a substantive sense the results are difficult to interpret, because Table 3B shows that the only stringency elasticity that is significantly different from zero (without as well as with correction for selection effects) refers to non-resource-based, pollution extensive industries. The theoretical expectation, however, is that resource-based, pollution intensive industries would be most severely affected by stringency of environmental policy, and non-resource-based, pollution extensive industries would hardly be affected (see also Mulatu et al. 2001).

5. Conclusions

Given the enormous productivity in academic research there is an increasing need for adequate tools to summarize the available empirical literature. Meta-analysis can be viewed as such a tool, and consists of a series of statistical and econometric techniques to analyse statistical summary indicators of empirical studies performed in the past. There are, however, a number of persistent methodological pitfalls that may be detrimental to the validity of meta-analysis. The most important are: biased sample selection and publication processes, heterogeneity among the studies contained in the meta-analysis, and dependence among the observed effect sizes in the meta-sample.

The heterogeneity of underlying studies is usually accounted for in meta-analyses in environmental economics. The independence requirement, however, is oftentimes ignored. This may cause estimators to be biased or inefficient, although one can argue that lacking independence is to some extent mitigated by allowing for random effects among studies. The requirement of a sample and publication selection process that is free of bias has received only fairly limited attention in environmental economics as well.

This paper has discussed a number of the available techniques to detect and even remedy publication bias. These techniques range from the use of sampling frames, via eyeball assessment of graphs, and univariate and bivariate statistics, to multivariate regression frameworks in which the selection process is modelled explicitly.

Most of these techniques to assess publication bias are illustrated by means of examples referring to price and income elasticities of residential water demand, and to stringency elasticities of intemational trade flows with respect to environmental policy. In most of the illustrations, publication bias-is detected in (almost) all samples, except for the so-called file drawer test that fails to detect publication bias. It is also of note that the most sophisticated technique, based on weighted distribution theory, is slightly more cumbersome to apply. In particular. overrepresentation of extremely smal1 p-values may lead to lack of convergence when applying the maximum likelihood routines (as in the

price elasticity example). For income elasticities we find compelling evidence that publication bias, caused by a publication screening process that favours positive income elasticities with small *p*-values, has a major impact on the results of the meta-analysis. The specific data constellation in the stringency elasticity example shows that a large sampling variance as compared to the conditional variance of the effect size may make the use of weighted distribution theory more difficult as well.

Notwithstanding the above, it is evident that publication bias is a serious issue that deserves proper attention in environmental economic meta-analyses. The fact that this area of research is still in development may make this more difficult (among other things because commercial software is not yet available). At the same time, however, this opens interesting vistas for new research. For economics it seems highly relevant to further investigate the consequences of multiple sampling of effect size estimates from the studies underlying the meta-analysis, because in general the number of studies available in economics is rather low. It also seems that further research on the applicability of the two-stage Heckman approach in meta-analysis is an attractive option, because this approach is not very well known outside economics.

References

Ashenfelter, O., C. Harmon and H. Oosterbeek, A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias, *Labour Economics,* vol. 6, 1999, pp. 453-470.

Begg, C.B., Publication Bias, in: H. Cooper and L.V. Hedges (eds.), *The Handbook of Research Synthesis*, Sage Foundation, New York, 1994.

Bergh, J.C.J.M. van den, K.J. Button, P. Nijkamp and G.C. Pepping, *Meta-Analysis in Environmental Economics,* Kluwer, Dordrecht, 1997.

Boyle, K.J., G.L. Poe and J.C. Bergstrom, What do we Know about Groundwater Values? Preliminary Indications from a Meta-Analysis of Contingent-Valuation Studies, *Americun Journal of Agricultural Economics,* vol. 76, 1994, pp. 10551061.

Brouwer, R., I.H. Langford, I.J. Bateman, T.C. Crowards and R.K. Turner, *A Meta-Analysis of Wetland Contingent Valuation Studies,* CSERGE Working Paper GEC 97-20, Centre for Social and Economic Research on the Global Environment (CSERGE), University of East Anglia and University College London, 1997.

Button, K., An Evaluation of the Potential of Meta-Analysis in Value Transfer and Function Transfer, in: R.J.G.M. Florax, P. Nijkamp and K. Willis (eds.), *Compururive Environmenrul Economic Assessmenr,* Edward Elgar, Cheltenham, 2001.

Card, D. and A. Krueger, Time-Series Minimum-Wage Studies: A Meta-Analysis, *American Economic Review,* vol. 85, 1995, pp.238-243.

Dalhuisen, J.M., R.J.G.M. Florax, H.L.F. de Groot and P. Nijkamp, *Price and Income Elusriciries of*

Residential Water Demand: Why Empirical Estìmates Dìffer, Research Paper, Tinbergen Institute, Amsterdam, 200 1.

Dear, K.B.G. and C.B. Begg, An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis, *Statistical Scìence,* vol. 7, 1992, pp. 237-245.

Glass, G.V., Primary, Secondary and Meta-Analysis of Research, *Educatìonal Research,* vol. 5, 1976, pp. 3-8.

Glass, G.V., B. McGaw and M.L. Smith, *Meta-Analysis in Social Research,* Sage Publications, Beverly Hills, 198 1.

Greene, W.H., *Econometric Analysis,* Prentice-Hall, Englewood Cliffs, 1993.

Greenhouse, J.B. and S. Iyengar, Sensitivity Analysis and Diagnostics, in: H. Cooper and L.V. Hedges, *The Handbook of Research Synthesis,* Sage Foundation, New York, 1994.

Heckman, J.J., Sample Selection Bias as a Specification Error, *Econometrica,* vol. 47, 1979, pp. 153-161.

Heckman, J., Varieties of Selection Bias, *American Economìc Review,* vol. 80, 1990, pp. 3 13-3 18.

Hedges, L.V., Estimation of Effect Size under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences, *Journal of Educatìonal Statistics,* vol. 9, 1984, pp. 61-85.

Hedges, L.V., Directions for Future Methodology, in: K.W. Wachter and M.L. Straf, *The Future of Meta-Analysis,* Russell Sage Foundation, New York, 1990.

Hedges, L.V., Modeling Publication Selection Effects in Meta-Analysis, *Statistical Scìence,* vol. 7, 1992, pp. 246-255.

Hedges, L.V., The Promise of Replication in Labour Economics, *Labour Economics,* vol. 4, 1997, pp. 111-114.

Hedges, L.V. and I. Olkin, *Statistical Methods for Meta-Analysis,* Academic Press, London, 1985.

Hunter, J.E. and F.L. Schmidt, *Methods of Meta-Analysis, Correctìng Error and Bìas in Research Findings,* Sage Publications, London, 1990.

Iyengar, S. and J.B. Greenhouse, Selection Models and the File Drawer Problem (with discussion), *Statistical Science,* vol. 3, 1988, pp. 109-135.

Lane, D.M. and W.P. Dunlap, W.P., Bias Resulting from the Significance Criterion in Editorial Decisions, *British Journal of Mathematica1 Statistic Psychology,* vol. 31, 1978, pp. 107-112.

Light, R.J. and D.B. Pillemer, *Summing Up: The Science of Reviewìng Research,* Harvard University Press, Cambridge, 1984.

Light, R.J. and P.V. Smith, Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies, *Harvard Educational Review,* vol. 41, 1971, pp. 429-471.

Mulatu, A., R.J.G.M. Florax, and C.A.A.M. Withagen, *Environmental Regulation and Competitiveness: A Meta-Analysis of International Trade Studies,* Research Paper, Tinbergen Institute, Amsterdam, 200 1.

Orwin, R.G., A Fail-Safe *N* for Effect Size in Meta-Analysis, *Journal of Educational Statistics*, vol. 8, 1983, pp. 157-159.

Petitti, D.B., *Meta-Analysis, Decision Analysis, and Cost Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine,* Oxford University Press, New York, 1994

Rosenthal, R., The "File Drawer Problem" and Tolerance for Null Results, *Psychological Bulletin,* vol. 86, 1979, pp. 638-461.

Rosenthal, R., An Evaluation of Procedures and Results, in: K.W. Wachter and M.L. Straf, *The Future of Meta-Analysis*, Russell Sage Foundation, New York, 1990.

Rosenthal, M.C., The Fugitive Literature, in: H. Cooper and L.V. Hedges (eds.), *The Handbook of Research Synthesis,* Sage Foundation, New York, 1994.

Schwartz, J., Air Pollution and Daily Mortality: A Review and a Meta-Analysis, *Environmental Economics,* vol. 64, 1994, pp. 36-52.

Smith, V.K., Can we Measure the Economic Value of Environmental Amenities?, *Southem Economic Journal,* vol. 56, 1989, pp. 865-878.

Smith, K.V. and J.C. Huang, Hedonic Models and Air Pollution: Twenty-Five Years and Counting, *Environmental and Resource Economics,* vol. 3, 1993, pp. 381-394.

Smith, K.V. and J.C. Huang, Can Hedonic Models Value Air-Quality? A Meta-Analysis of Hedonic Property Value Models, *Journal of Political Economy, vol.* 103, 1995, pp. 209-227.

Smith, V.K. and Y. Kaoru, Signals or Noise: Explaining the Variation in Recreation Benefits Estimates, *American Journal of Agricultural Economics,* vol. 72, 1990, pp. 419-433.

Smith, V.K. and L. Osbome, Do Contingent Valuation Estimates Pass a "Scope" Test? A Meta-Analysis, *Journal of Environmental Economics and Management,* vol. 3 1, 1996, pp. 287-301.

Sturtevant, L.A., F.R. Johnson and W.H. Desvouges, *A Meta-Analysis of recreational Fishing*, Triangle Economic Research, Durham, 1995.

Vevea, J.L. and L.V. Hedges, A Genera1 Linear Model for Estimating Effect Size in the Presence of Publication Bias, *Psychometrika,* vol. 60, 1995, pp. 419435.

Walsh, R.G., D.M. Johnson and J.R. McKean, Market Values from Two Decades of Research on Recreational Demand, in: A.N. Link and V.K. Smith (eds.), *Advances in Applied Economics Volume* 5, JAI Press, Greenwich, 1989.

Woodward, R.T. and Y.-S. Wui, The Economic Value of Wetland Services: A Meta-Analysis, *Ecological Economics,* **vol.** 37, 2001, pp. 257-270.

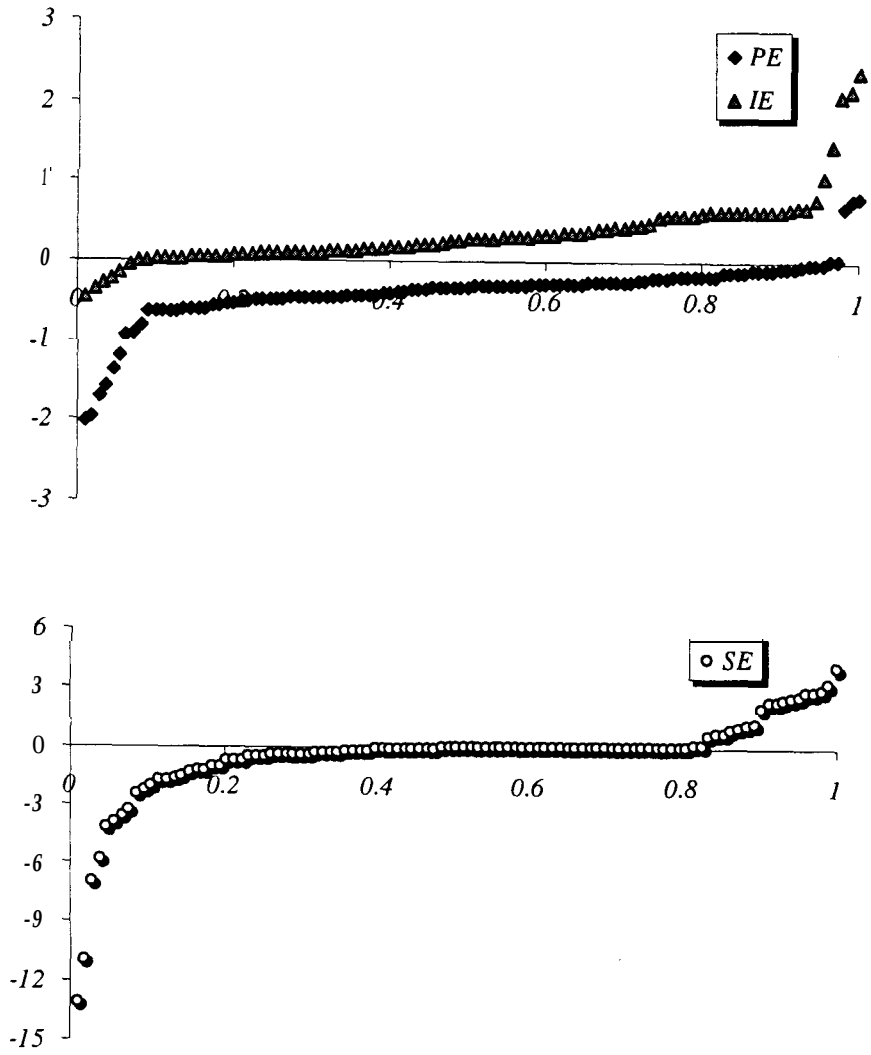

FIGURE 1. Price (*PE*), income (*IE*) and stringency (*SE*) elasticities, ordered according to magnitude in deciles of the meta-samples.

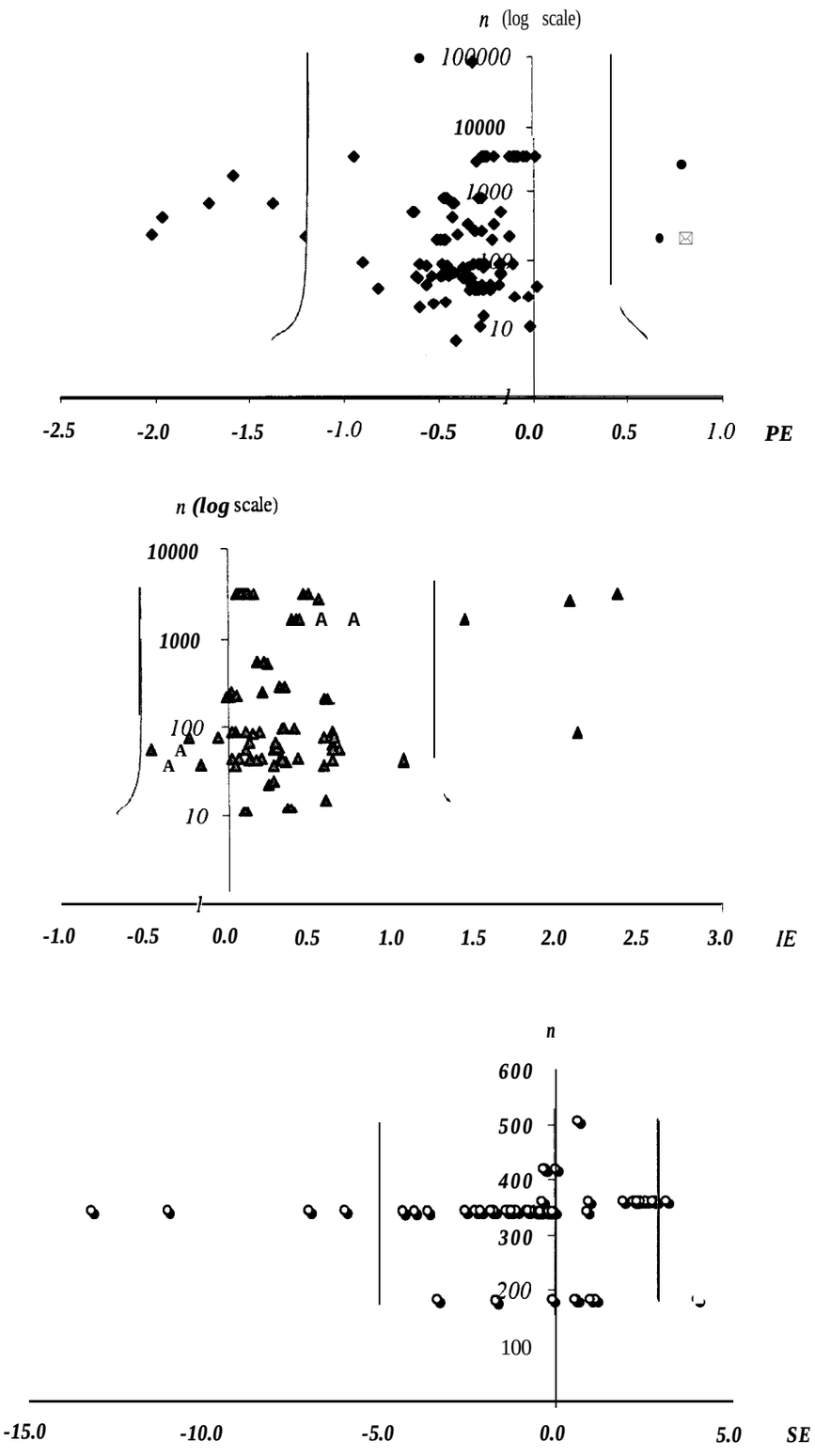

FIGURE 2. Funnel graphs for price (*PE*), income (*IE*) and stringency (SE) elasticities.

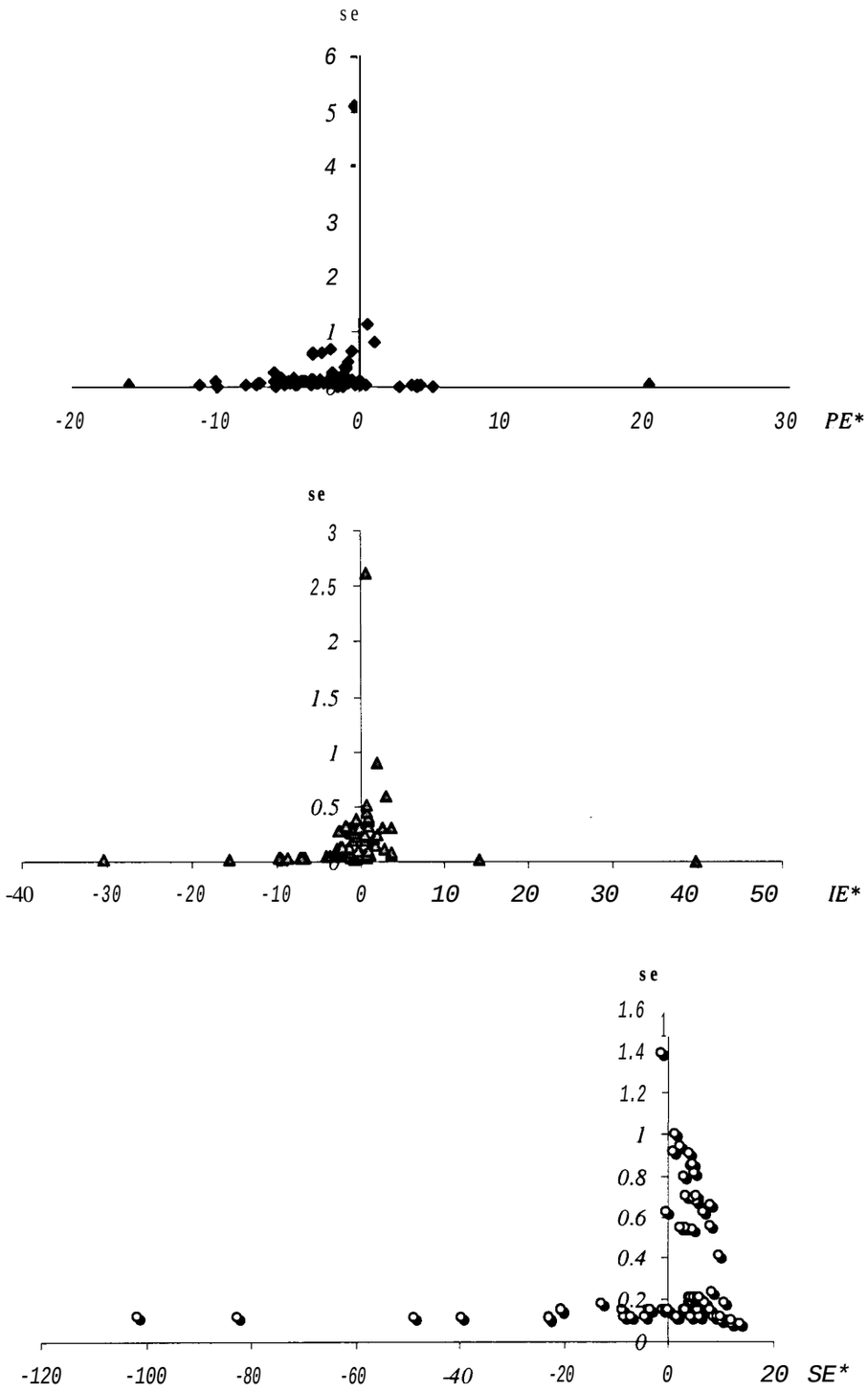

FIGURE 3. Graphs of standardized price *(PE*)*, income *(lE*)* and stringency (SE*) elasticities against their estimated standard errors.

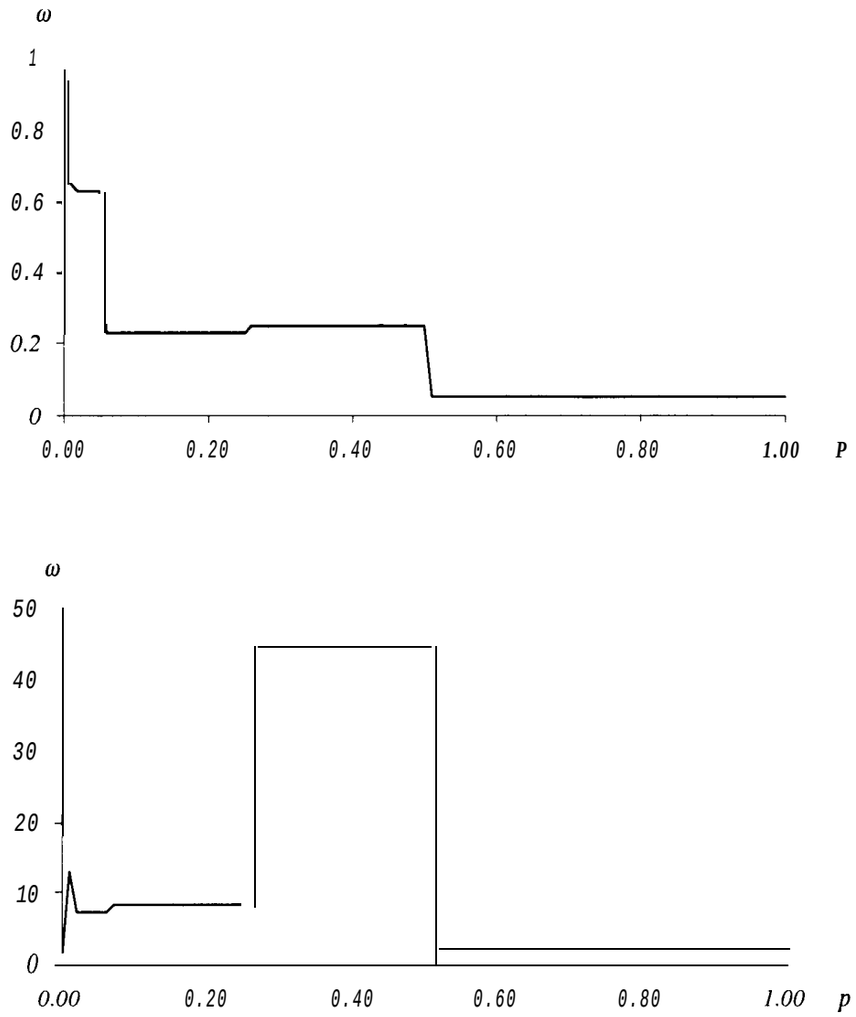

FIGURE 4. Step functions of estimated weights for different p-intervals of income (top) and stringency (bottom) elasticities.

TABLE 1. A Pearson χ^2 test for the difference between observed and expected p-values categorized in six discrete intervals for price, income, and stringency elasticities.[a]

Interval (max)	*Price elasticities*			*Income elasticities*			*Stringency elasticities*		
	Observed	*Expected*	χ^2	*Observed*	*Expected*	χ^2	*Observed*	*Expected*	χ^2
0.001	61	23.11	62.12***	39	36.02	0.25	24	45.69	10.30***
0.01	17	11.99	2.09	9	6.73	0.77	10	3.32	13.46***
0.05	13	16.95	0.92	13	7.76	3.54*	5	3.15	1.08
0.25	12	28.46	9.52***	11	12.87	0.27	8	4.69	2.34
0.50	2	15.55	11.81***	11	8.37	0.82	29	3.27	202.13***
1.00	5	13.93	5.73**	7	18.26	6.94**	27	42.88	5.88**
Pearson			92.18***			12.59**			235.18***

[a] Significance is indicated by ***, ** and * for the 0.01, 0.05 and 0.10 level.

TABLE 2A. Regression results for models with and without predictors and publication bias for income elasticities of residential water demand.[a]

	Without predictors			With predictors		
	no p-values	p-values	included	no p-values	p-values	included
Constant	0.27***	0.02		0.26***	0.01	
	(0.02)	(0.12)		(0.03)	(0.13)	
***Marginal** price*				-0.01	0.005	
				(0.06)	(0.10)	
Shin price				0.43**	0.69**	
				(0.19)	(0.29)	
***Difference** variable*				-0.11	-0.15	
				(0.10)	(0.15)	
Discrete/continuous				0.35***	0.46**	
				(0.13)	(0.20)	
$p = 0.001$		1.00 fixed[b]			1.00 fixed[b]	
$p = 0.01$		0.65**			0.62**	
		(0.28)			(0.26)	
$p = 0.05$		0.63**			0.61**	
		(0.27)			(0.26)	
$p = 0.25$		0.23*			0.23*	
		(0.12)			(0.12)	
$p = 0.50$		0.25*			0.26*	
		(0.15)			(0.16)	
$p = 1.00$		0.05			0.05	
		(0.04)			(0.04)	
***Variance** component*	0.03***	0.07***		0.03***	0.07***	
	(0.01)	(0.02)		(0.01)	(0.02)	
Log-likelihood	30.16	13.18		17.74	1.51	
LR for selection	16.99***			16.23***		

[a] Estimated parameters are given with estimated standard errors in Parentheses. Significance is indicated with ***, ** and * for the 0.01, 0.05 and 0.10 level.
[b] Weight exogenously fixed at unity.

TABLE 2B. Estimated means for income elasticities of residential water demand with and without adjustment for publication bias."

Price type	*Condition*	*Without selection*		*With selection*	
		Mean	*Standard error*	***Mean***	*Standard error*
Average	*None*	0.26***	0.03	0.01	0.13
	***Difference** variable*	0.15	0.11	-0.14	0.20
	Discrete/continuous	0.61***	0.14	0.47**	0.22
Marginal	*None*	0.25***	0.05	*0.01*	0.14
	***Difference** variable*	0.15*	0.09	-0.14	0.17
	Discrete/continuous	0.60***	0.14	0.48**	0.22
Shin	*None*	0.69***	0.18	0.70***	0.26
	***Difference** variable*	0.58***	0.21	0.55*	0.30
	Discrete/continuous	1.04***	0.23	1.16***	0.33

[a] Predicted means and standard errors of the predicted means are presented. Significance for a two-sided test of the mean being different from zero is indicated with ***, ** and * for the 0.01, 0.05 and 0.10 level.

TABLE 3A. Regression results for models with and without predictors and publication bias for stringency elasticities of international trade flows.[a]

	Without predictors		With predictors	
	no p-values	p-values included	no p-values	p-values included
Constant	-0.49**	-1.54**	-1.68***	-3.17***
	(0.23)	(0.70)	(0.44)	(0.74)
Pollution intensive			2.02***	3.37***
			(0.66)	(1.07)
Resource-based			1.37**	1.91**
			(0.55)	(0.82)
Interaction[b]			-1.89**	-2.19
			(0.94)	(1.66)
$p = 0.001$		1.00		1.00
		fixed[c]		fixed[c]
$p = 0.01$		12.76**		13.99**
		(5.60)		(6.24)
$p = 0.05$		7.19*		7.71*
		(3.94)		(4.26)
$p = 0.25$		8.23**		8.58**
		(3.99)		(4.20)
$p = 0.50$		44.38**		45.02**
		(17.50)		(17.95)
$p = 1.00$		2.29**		2.34*
		(1.15)		(1.20)
Variance component	5.16***	8.95***	4.67***	7.47***
	(0.74)	(1.77)	(0.67)	(1.44)
Log-likelihood	463.90	333.59	453.73	322.21
LR for selection	130.31***		131.51***	

[a] Estimated parameters are given with estimated standard errors in parentheses. Significance is indicated with ***, ** and * for the 0.01, 0.05 and 0.10 level.
[b] Pollution intensive x Resource-based
[c] Weight exogenously fixed at unity.

TABLE 3B. Estimated means for stringency elasticities of international trade flows with and without adjustment for publication bias."

Price type	Condition regarding pollution	Without selection Mean	Without selection Standard error	With selection Mean	With selection Standard error
Resource-based	Intensive	-0.18	0.59	-0.08	1.27
	Extensive	-0.31	0.33	-1.26	0.77
Non-resource-based	Intensive	0.35	0.49	0.20	0.99
	Extensive	-1.68***	0.44	-3.17***	0.74

[a] Predicted means and standard errors of the predicted means are presented. Significance for a two-sided test of the mean being different from zero is indicated with ***, ** and * for the 0.01, 0.05 and 0.10 level.