April, 2018

Skating on Thin Evidence:
Implications for Public Policy

Hristos Doucouliagos,[1] Martin Paldam,[2] and T.D. Stanley[3]

European Journal of Political Economy, 54, 16-25, 2018

Abstract: Good public policy needs to be evidence based. However, the evidence base is thin for many policy issues. How can policy makers best respond to such thin areas of research that are also quite likely to change over time? Our survey investigates the evolution of the econometric evidence base for 101 economic issues, drawing upon 42,578 effect sizes (mainly elasticities and correlations) from 4,300 econometric studies. We evaluate the performance of six approaches to early research assessment: the simple unweighted mean; the median; the Paldam, "divide by 2" rule of thumb; the unrestricted weighted least squares (*WLS*) weighted average; the *PET-PEESE* meta-regression correction for publication bias; the weighted average of the adequately powered (*WAAP*); and *WAAP-WLS.* Lowest prediction errors are found in the Paldam rule of thumb and *WLS. WLS* typically reduces the initial exaggeration of thin evidence by half.

Keywords: thin evidence, meta-regression, *WLS, WAAP*, Paldam rule of thumb
JEL codes: C1, H00, H5

[1] Corresponding author. **Doucouliagos**: Department of Economics, Alfred Deakin Institute, and Deakin Lab for Meta-Analysis of Research, Deakin University, and Institute for the Study of Labor (IZA). Email: douc@deakin.edu.au.
[2] **Paldam**: Department of Economics and Business, Fuglesangs Allé 4. University of Aarhus, 8210 Aarhus V, Denmark. E-mail: mpaldam@econ.au.dk.
[3] **Stanley**: Deakin Lab for Meta-Analysis of Research, Deakin University and Hendrix College. Email: Stanley@hendrix.edu.

1. Introduction

Policy makers frequently refer to the need for evidence-based policy.[4] This requires credible research. However, concerns are mounting that the evidence base in many of the social sciences lacks credibility. For example, in economics and business research, there is widespread failure to replicate empirical findings (Dewald, Thursby and Anderson, 1986; Hubbard and Vetter, 1992; Evanschitzky and Armstrong; 2010; Camerer *et al*. 2016; and Miguel and Christensen, 2017), most studies lack adequate statistical power (Ioannidis, Stanley and Doucouliagos, 2017), much of the evidence base is afflicted with publication selection bias (Doucouliagos and Stanley, 2013)[5], heterogeneity in reported findings is excessive in almost all fields, and there is also some evidence of outright fraud (Bailey *et al*., 2001). Needless to say, these are serious issues that may obscure the real research message and are, in one way or another, related to insufficient data.

The focus of this paper is on an issue that have received little attention—thin evidence. By this, we mean an evidence base that is either just emerging or is still too small to estimate the key policy parameter, reliably. For example, there are only a handful of empirical studies from which public policy can be informed for newly emerging policy issues such as: the effectiveness of quantitative easing or the effectiveness of fat taxes.[6] Moreover, evidence is particularly thin for most specific countries. Researcher incentives and career concerns are such that more econometric evidence will be produced for larger, more developed countries, notably the US and other OECD countries rather than Central Asia and Africa.

To assess the importance of thin evidence, we survey 101 areas of economics research containing 42,578 estimated effects from 4,300 econometric studies. Our survey finds that all summary estimates from thin evidence poorly predict how a mature research base evolves. However, we also find that a simple rule, divide by two, and a simple weighted least squares (WLS) weighted average do much to reduce the bias and exaggeration of early research studies. We believe that this is the first empirical study of thin evidence, but several Monte Carlo simulations studies exist (Moreno et al., 2009; Stanley and Doucouliagos, 2105; Stanley et al., 2017; Stanley, 2017). These simulation studies investigate conventional statistical properties

[4] Examples include U.K. Cabinet Office (1999) and the Office of Management and Budget (2012). The 3ie/DFID initiative for the systematic review of international development research is all about evidence based policy http://www.3ieimpact.org/.

[5] As of March 2018 Google scholar has 3.6 million hits to the term 'publication bias'. This is a problem for all fields of research.

[6] While the discussion in the paper is on public policy, decision makers in the private sector encounter the same issues when they need to make decisions based on the extant thin evidence base.

such as: bias, mean squared error (MSE), and coverage of meta-analysis estimators for as few as 5 or 10 research estimates. However, they also caution against relying on meta-analysis when so little research is available.

The paper is set out as follows. Section 2 outlines the issues at stake. Section 3 presents the meta-analysis methods used in this survey. Section 4 discusses the data. Section 5 presents and discusses the results for predictive accuracy, while Section 6 presents and discusses the results for research exaggeration. Section 7 concludes the paper.

2. Skating on thin evidence

The problem of a thin evidence base is highlighted in Table 1, which reports the median number of years it takes for empirical economics to produce 5, 10, 20, 40, and 80 studies on a given topic.[7] On average, it takes 7 years for a literature to produce and publish 5 empirical studies and 15 years for 20 studies to be made available. It then takes more than two decades, on average, for a literature to report 40 empirical studies. In the meantime, decisions must be made. Do policy makers need to wait for research to mature, or can useful inferences be made even when the evidence base is thin?

Table 1. Median number of years for empirical studies to be publicly available

First k studies:	(1) $k = 5$	(2) $k = 10$	(3) $k = 20$	(4) $k = 40$	(5) $k = 80$
Median number of years	7	10	15	22	31
Median number of	23	53	144	333	523
Number of literatures	101	92	73	41	9

Source: Authors' calculations based on the 101 research literatures listed in Appendix A.

The problem upon which this paper focuses is illustrated by the literature on the effect of development aid on economic growth—see Figure 1. The first three panels illustrate the distribution of reported estimates for the first 5, 10, and 20 studies, respectively. The last panel uses all reported data as of 2015. As research deepens over time, the average correlation of aid effectiveness falls from 0.28 in the first 5 studies to 0.07 after all 143 studies are conducted.

[7] The unit of analysis in this paper is studies rather than estimates reported within studies.

As a second example, Figure 2 illustrates the evolution of the Rose effect - the trade effect of a currency union - literature. The inference from the first 5 studies is that there is a large currency effect, with the mean value of gamma value of 1.07.[8] However, the mean value of gamma falls to 0.51 as the literature matures.

Figure 1. The evolution of the aid effectiveness literature

Figure 1a. For $k = 5$, $N = 26$

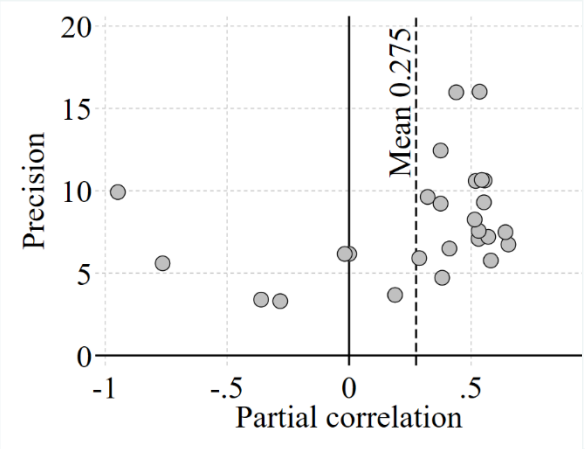


Figure 1b. For $k = 10$, $N = 67$

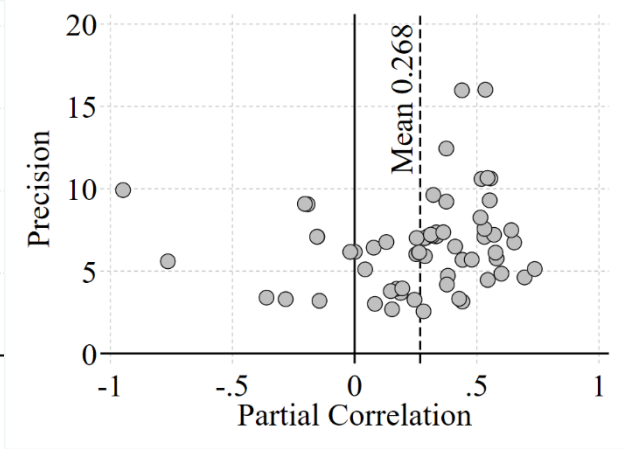


Figure 1c. For $k = 20$, $N = 171$

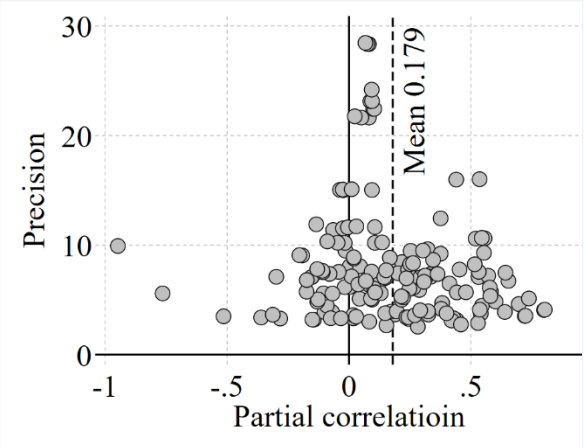
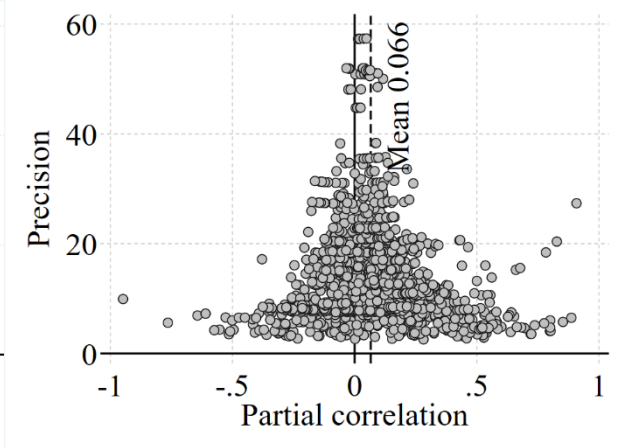


Figure 1d. For $k = 139$, $N = 1{,}777$



Source: Constructed from data reported by Doucouliagos and Paldam (2015). *N* is the number of estimates, *k* is the number of papers.

[8] The trade effects of a currency union are typically estimated by regressing the logarithm of trade between nations on a range of variables, one of which is a dummy variable taking the value of 1 if trading partners are in a currency union. Gamma is the estimated coefficient on this dummy variable. The mean values in Figure 2 are unweighted and not corrected for publication bias.

Figure 2. The evolution of the literature on the Rose effect

Figure 2a. For $k = 5$

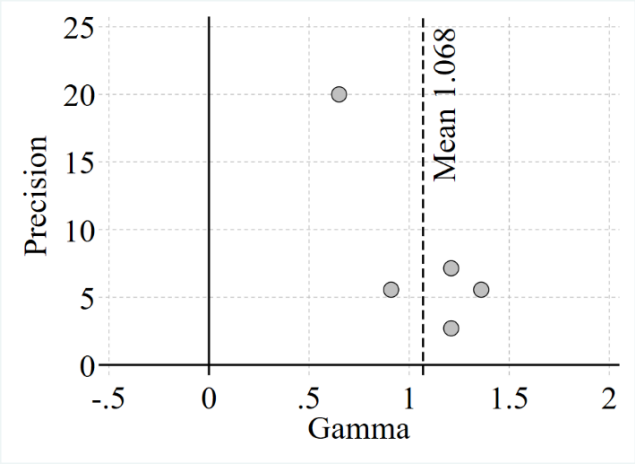


Figure 2b. For $k = 10$

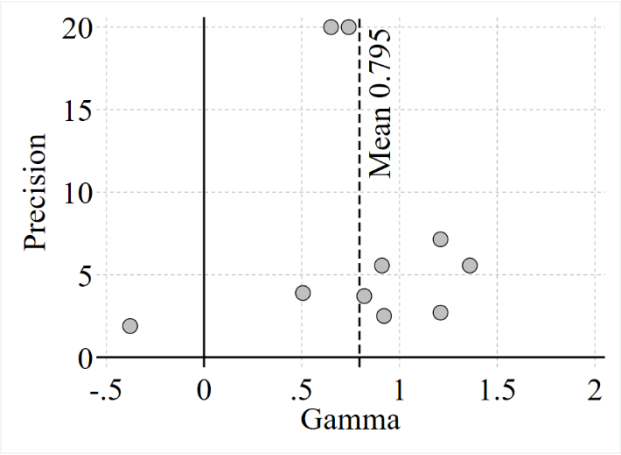


Figure 2a. For $k = 20$

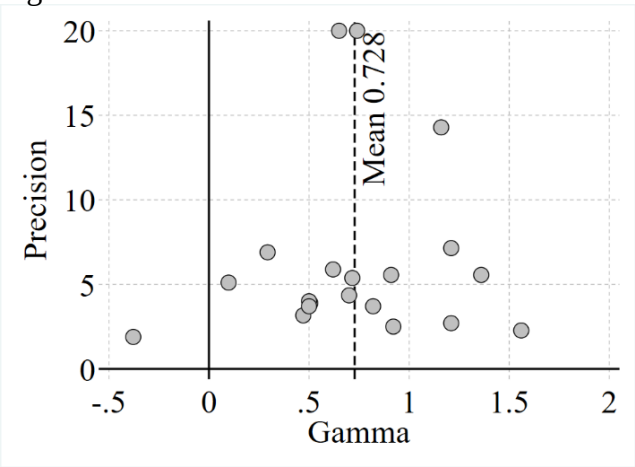
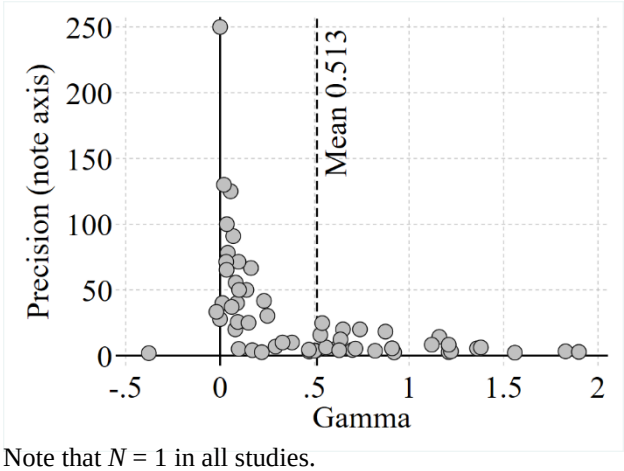


Figure 2b. For all $k = 61$



Source: Constructed from data reported by Havranek (2010). Note that $N = 1$ in all studies.

Over time the evidence base grows when there is sufficient interest in a topic, or it stagnates if there are insufficient incentives for researchers to supply new parameter estimates in a given area. The problem for policy makers is that the reported parameter estimates may change over time. These changes can reflect one of four scenarios. First, there may be genuine changes in the underlying 'true' effect. That is, the 'true' effect might be time-varying. Second, with a thin evidence base, there is a higher risk that one study might bear an undue influence on the inferences that a literature appears to be drawing. With only a handful of studies, it is possible that one or two studies capitalize on sampling error or some random bias. Consequently, these studies can exert undue influence or leverage on an emerging area of research. However, if there is a persistent pattern of exaggeration across many areas of research, then something more than sampling error or random bias is at play. Third, researchers and/or journals may prefer to

publish inflated effects. This has been associated with the winner's curse, where "the more extreme, spectacular results (the largest treatment effects, the strongest associations, . . .) may be preferentially published" (Young et al., 2008, p. 1418). Perhaps initially, completion for publication reinforces such exaggeration that might lessen over time. Fourth, it has been suggested that most extreme and contradictory results might be published very early rather the later. This is known as the 'Proteus phenomenon', whereby the earlier literature has excessive variation that moderates over time (Ioannidis and Trikalinos, 2005). These effects may give the *appearance* of policy effectiveness larger than they really are.

Though publication bias in the form of exaggeration of results seems to be common (Ioannidis et al., 2017), it is unclear if it decreases or increases over time. First, data typically grows over time, both because many organizations make data more accessible and because time periods accumulate. This may reduce the need to exaggerate as higher statistical power is more likely to reveal statistically significant effects. In addition, it decreases the variability of results. Second, greater computer power and user-friendliness of statistical packages reduces the costs of running regressions. This increases the pool of results that can be searched and reduces the effort required to do so, potentially exaggerating findings further (Paldam, 2017).

In this paper, we are interested in the ability to draw reliable inferences from the first few studies, e.g. from the first 5 or 10 studies. What methods should policy makers adopt to summarize the parameter estimates from a thin evidence base? Or, is it ever wise to base policy on thin evidence? Can meta-analysis offer tools for drawing better inferences from a thin evidence base?

In one sense, meta-analysis was designed to tackle the problem of thin evidence. By pooling estimates from different studies, weak and underpowered empirical studies are combined and this thereby increases statistical power. However, little is known about the relative performance of meta-regression methods when studies are rare and evidence is thin. Hence, a second motivation for this paper is to assess the performance of recently developed meta-regression methods as a research base evolves.

3. Meta-analysis methods

"(T)he data almost never speak for themselves, especially when there are missing data."
–Heckman (2000:3)

We consider the evolution of an empirical literature at six distinct stages: 5 studies, 10, 20, 40, 80 and, lastly, all reported empirical studies. Though arbitrary, these stages are broadly indicative of the evolution of empirical literatures. A literature with only 5 empirical studies is underdeveloped, the evidence base is very thin and quite likely to evolve in one direction or another. A literature with 10 or even 20 studies will have experienced some development but it is still relatively thin. At 40 studies the literature is maturing. A literature with 80 studies can be considered to be sufficiently large and mature to be taken seriously.

Regardless, policy makers need to base their decisions upon some parameter estimates from this evolving evidence base. Often, the most pressing policy issues are those that are newly emerging and for which, by definition, the evidence base is thin. Thin evidence is particularly problematic for observational research. Five or ten *large* randomized clinical trials (RCTs) might be sufficient to draw valid inferences on the effectiveness of a new drug when carefully conducted and fully reported. However, in general, the same number of econometric studies of observational data is unlikely to be sufficient. Observational data are by their nature less reliable and more vulnerable to excess heterogeneity, misspecification, and selection bias than are large RCTs.

Our interest is on identifying which of a set of established and newly developed meta-regression analysis estimators (discussed below) is best suited to provide usable estimates when the literature is relatively thin, i.e. when there are only 5 or 10 reported studies. We are particularly interested in the performance of these estimators when the evidence base is newborn, i.e. when there are only 5 available empirical studies. We evaluate the performance of six estimators of the average value of the parameter of policy interest, against four proxies for the underlying 'true' effect.

3.1 Estimators

We compare the performance of six meta-analysis averages in tracking the evolution of a literature.

Unweighted mean: is the simple average of the first 5, 10, 20, …, studies. No correction is made for publication selection bias nor are the better, more precise, estimates given greater weight.[9]

[9] Many government departments carry out literature reviews of the evidence base and merely calculate simple unweighted averages. There are exceptions however. For example, the U.S. EPA uses weighted averages derived from meta-regression analysis when reviewing parameters such as the value of a statistical life (Viscusi and Aldy, 2003).

Median: is the 50% percentile and is quite robust to outliers. The *median* is also not corrected for publication bias nor is it weighted.

Paldam rule of thumb: divides the unweighted mean by two.[10] No further corrections are made.

WLS: is an unrestricted weighted least squares weighted average that has recently been shown to possess superior properties (smaller bias and mean squared error) relative to conventional random-effects meta-analysis when there is publication bias (Stanley and Doucouliagos, 2015; 2017).[11] There is no explicit correction for publication selection but *WLS* will passively reduce publication bias when it is present at no practical cost if it is not. *WLS* uses the same inverse variance weights and gives the exact same point estimate as does conventional fixed-effect meta-analysis. However, *WLS* does not assume a fixed effect and automatically accommodates any heterogeneity found in the research record.

PET-PEESE: like *WLS*, uses inverse variance weights; however, this method also corrects the evidence base for publication selection bias. This is a conditional estimator (Stanley and Doucouliagos, 2014). Initially, the *FAT-PET* meta-regression of the estimated effect sizes is run on a constant and the estimated standard errors of these effect sizes. The coefficient on the constant provides an estimate of the effect size corrected for publication selection bias. If there is a statistically significant effect after correcting for selection bias,[12] then a non-linear term, the standard errors squared, becomes the meta-regression's independent variable. See Stanley and Doucouliagos (2012; 2014) for greater detail.

WAAP-WLS: is the *WLS* weighted average calculated on only those estimates that have adequate statistical power (Ioannidis *et al*., 2017).[13] When there are no estimates with sufficient statistical power, *WAAP* cannot be calculated and *WLS* estimate is used in its place (Stanley *et al*., 2017). *WAAP* cannot be used for many of the research literatures in our database especially when the evidence is very thin. In contrast, *WAAP-WLS* can be applied to all of them.

[10] The Paldam rule of thumb is attributed to Martin Paldam. It emerged from observations of the results of a number of meta-studies. Martin Paldam suggested that whenever anyone did not know much about a specific empirical literature, a simple approach would be to just divide authors' estimated coefficients by 2. The idea behind this is that authors often seem to exaggerate their results, making them appear to be 'too good' to be true.
[11] In meta-analysis, the terms fixed- and random-effects denote the specific weights used to pool estimates from various studies and not to the panel structure of the data; for details see Stanley and Doucouliagos (2012).
[12] This involves a test of the null hypothesis that the mean effect is zero. Because multiple estimates per study are quite common, we correct the standard errors of the meta-regressions for clustering of estimates within studies.
[13] Following (Cohen, 1988) and Ioannidis *et al.* (2017), an estimate has adequate power if its standard error is small enough to ensure that there is a 80% probability of finding a true effect equal to the *WLS* estimate. This requires the standard error to be smaller than the absolute value of the unrestricted weighted least squares weighted average divided by 2.8. 2.8 is the sum of the conventional critical value, 1.96, and the 80% threshold for the cumulative normal distribution (0.84) that defines the conventional 80% power requirement (Cohen, 1988).

3.2 'True' effect

In order to evaluate the above estimators, we need a measure of the underlying 'true' effect. Of course, we cannot know what this is. We can only estimate it. Here, we use four proxies of the 'true' effect: (a) *WLS*; (b) *PET-PEESE*; (c) *WAAP*; and (d) *WAAP-WLS*, calculated from *all the available estimates* in the research base. The difference between the estimators discussed in 3.1 above and these proxies for 'true' effect is that the former use the data from only the first 5, 10, or 20 studies, whereas the later use *all* relevant research studies.

Here, we do not allow the 'true' effect to evolve over space (e.g. countries) or time. We assume that the mean 'true' effect is fixed over time and that it is best approximated using all available research. This assumption is a limitation of our survey. However, it is not clear whether these observed declines in effect sizes reflect genuine changes in the underlying 'true' effect, or whether they are artefacts of research design, changes in research practices or publication bias (Ioannidis and Trikalinos, 2005). No matter how rigorous our methods may become, there will remain some ambiguity about whether the underlying phenomenon is truly declining over time or whether it is econometric methods and measures that identify smaller effects. For the purposes of practical policy, it is largely immaterial whether the 'true' effectiveness of a particular policy effect is declining or whether our best research is saying that it is decreasing. In either case, policy makers are likely to find the consequences of their chosen policy intervention is less than what they had hoped for when planning this intervention, years before.

To recap, the *WLS* involves running the following WLS regression that uses inverse variance weights (1/ SE^2):

(1) $r_{ij} = \beta_0 + v_{ij}$

where r_{ij} denotes a measure of the effect size (e.g., partial correlation or elasticity), i and j index the ith estimate from study j, and the *v* are random errors.

PET-PEESE involves first running the *FAT-PET* regression:

(2) $r_{ij} = \beta_0 + \beta_{ij} SE_{ij+} v_{ij}$

where *SE* denotes the standard error of the effect size. If Eqn. (2) suggests that there is a genuine effect net of publication selection bias (i.e. β_0 is statistically significant), then the *PET-PEESE* regression is run:

(3) $r_{ij} = \beta_0 + \beta_{ij} SE_{ij}^2 + v_{ij}$

where SE^2 denotes the standard error squared, and all of these regression need to use a WLS routine with weights (1/ SE^2).

WAAP-WLS involves running Eqn. (1) on only those estimates that are adequately power. If there are no adequately powered estimated, then this conditional estimator uses the WLS estimate.

For all models (*WLS, PET-PEESE, WAAP-WLS*) β_0 provides an estimate of the average effect for 5, 10, 20 … studies and also the estimated WLS 'true' effect when all studies are used to estimate Eqn. (1).

3.3 Performance

We use two general criteria to assess the performance of each of the above six estimators relative to the four estimates of the 'true' effect: predictive accuracy and research inflation or exaggeration.

Mean absolute percentage error (MAPE) is commonly used in forecasting and in meta-analysis for predictive accuracy (Makridakis, Wheelwright and McGee, 1978; Rosenberger and Stanley, 2006). This is calculated as: $MAPE = \sum_{i=1}^{n} \left| \frac{estimate_i - 'true' effect}{'true' effect} \right| \cdot 100 / n$.

The benefit of MAPE is that it is scale independent and can be compared across the different datasets in our survey. However, MAPE may be biased in some situations. Hence, as part of robustness, we also consider the symmetric mean absolute percentage error (SMAPE). This involves a slight adjustment to the MAPE that uses the sum of the absolute values of the estimate and the 'true' effect in the denominator.

Our second performance measure is 'research inflation' or 'research exaggeration'. Ioannidis *et al.* (2017) define research inflation as the systematic exaggeration of the average reported estimate relative to the 'true' effect. In addition, we also assess the proportion of research areas where the early literature reports an effect in the opposite direction as the subsequent empirical record.

4. Data

The data used in this survey are drawn primarily from the meta-data used in a recent survey of power and bias in economics (Ioannidis *et al.*, 2017). In addition, we have included six studies not included in Ioannidis *et al.* (2017). In total, we use data from 101 meta-analyses that have documented the population of comparable empirical studies on a particular topic. We use the data collected by each of the 101 meta-analysis. That is, we apply the alternate estimators and

estimates of the 'true' effect to the data used by the 101 meta-analysis.[14] Assuming that these 4,300 studies and 42,578 estimates are representative, we can explore the evolution of empirical economics. Most of these 4,300 studies report multiple estimates, rather than a single estimate per study. Multiple estimates are preferred because selecting the 'best estimate' only increases bias–see Viscusi (2017).

Different effect size measures are used in different research areas. Specifically, the data consists of studies that use elasticities, correlations, and other measures (dollar values, standardized scores, etc.). The meta-studies are referenced in the online Appendix A.

Using data from prior meta-analyses has several advantages. Meta-analyses provide a feasible population of comparable parameter estimates on a specific research issue. This means that we do not have to read and code thousands of research papers; a herculean and prohibitive task. Having other researchers collect and code the relevant studies means that we can rely on experts on these specific topics, who better understand the nuances of a given research field.

We make no claim that the data are representative of empirical economics. The data are, however, representative of meta-analyses in economics. That is, they are a sample of typical areas of research that have been reviewed using the tools of meta-analysis. There are, obviously, many other areas that policy makers are interested in and which have either not been assessed using meta-analysis or where meta-analysis has been conducted but the data were not available to us.

First, we use the data as supplied by authors to estimate meta-averages. Second, we remove outliers from the data, even if the authors used data with outliers in their analysis. This involves running a *FAT-PET* regression on all observations and then identifying any observations with a standardized residual greater than 3.5 (in absolute value). This allows the removal of possible outliers from the data that might actually be errors in coding, either from the primary authors or somehow introduced by the meta-analysts.

Table 2 reports the distribution of the fields covered in our survey, according to *Journal of Economic Literature* classification category. Our sample is dominated by codes D, J, M, and O in terms of the number of meta-analyses and codes E, F, J, and O in terms of the number of econometric studies.

[14] This type of analysis is sometimes known as meta-meta-analysis or meta-research.

Table 2. Fields covered and some counts of the sample

	Journal of Economic Literature category	Number		
		Meta studies	Primary papers	Estimates
D	Microeconomics	12	466	1,604
E	Macroeconomics	7	451	8,045
F	International Economics	7	241	6,273
H	Public Economics	8	349	3,859
I	Health, Education, & Welfare	5	216	2,672
J	Labor & Demographic Economics	26	932	8,672
M	Business Administration & Business Economics	12	597	1,676
O	Development, Innovation, Technology, and Growth	18	773	7,831
Q	Agricultural & Natural Resource Economics	6	275	1,946
	All	101	4,300	42,578

As noted in the introduction, we are concerned about time variation in effect sizes. Table 3 reports the number of fields for which the estimated effect is falling (rising) and the median decline (rise) over time. The effect is falling for the great majority, 73%, of the literatures. Row 2 of Table 3 compares the change in the unweighted average between the value when all observations are used and when only 5 studies are available. The median decline in the effect size from inception of a research agenda to the time the meta-analysis was conducted is -39%.

Table 3. Evolution of effect sizes in empirical economics

	Falling over time	Rising over time
Number of research fields	74	27
Unweighted average median % change	-39%	40%

4.1 Is research growing faster?

The earliest research literature in our survey commenced in 1940 and the most recent began in 2008. Table 4 compares the growth in the younger literatures against the older literatures in terms of growth rates, splitting the data according to the average starting year, 1984. There appears to be a shift over time with newer literatures growing faster, perhaps due to the great expansion in the number of journals, and greater availability of data and increased computational power over recent decades. Nevertheless, even in the newer literatures it takes about 9 years to establish 10 empirical studies, which remains a rather long period.

Table 4. Median number of years for studies to be publicly available

First *k* studies:	(1) $k = 5$	(2) $k = 10$	(3) $k = 20$	(4) $k = 40$	(5) $k = 80$
Median number of years, all	7	10	15	22	31 (9)
Median number of years,	9 (43)	13	19	32	33 (8)
Median number of years,	6 (58)	9 (49)	12	13	25 (1)

Note: Older and younger are relative to 1984. Figures in brackets report the number of meta-studies or economic issues.
Source: Authors' calculations based on the 101 research literatures listed in Appendix A.

5. Predictive accuracy

For our analysis, we first sort the data in chronological order to represent the historical evolution of each research area. Next, as discussed above, we identify and remove outliers. Finally, we calculate bias, research inflation, and MAPE (and SMAPE), using the different estimators and alternate measure of the 'true' effect. Table 5 reports results for MAPE, for the first 5, 10, and 20 studies reported in the 100 empirical literatures. The rows list the six estimators calculated from the early stages of research and the columns list the four proxies for 'true' effect calculated from all estimates available at the time the meta-analyses were conducted. The last column reports the median MAPEs across the four 'true' effect measures, while the last row reports the median MAPEs across the six estimators for each of the four proxies of 'true' effect.

Table 5 reveals that the simple averages, Mean and Median, consistently give poor predictions of future findings when a research literature is in its infancy or its early stages. Surprisingly, the Paldam rule gives the lowest MAPEs (for k = 5 and 10). This is based on the median MAPE across the four methods for estimating 'true' effect. However, both *WLS* and *WAAP-WLS* have comparable and sometimes superior prediction accuracy and, along with Paldam, are distinctly better than *PET-PEESE*. *WLS* emerges as the clearly preferred early assessment method if SMAPE (symmetric mean absolute percentage error), rather than MAPE, is used to measure prediction accuracy—see Appendix B. Among meta-analysis methods, *PET-PEESE* is the worst performer; yet, it is the only method that attempts to correct the evidence base for publication selection. However, it should come as no surprise that a *regression* bias correction does not work well with only 5 or 10 studies (Stanley and Doucouliagos, 2017; Stanley, 2017; Stanley, Doucouliagos and Ioannidis, 2017). If we focus on the two best proxies of 'true' effect (*WLS* and *WAAP*), as determined by minimizing

prediction errors, the Paldam rule and WLS are essentially equivalent and the best, closely followed by *WAAP-WLS*.

Table 5. Median MAPE reported as a percent for k = 5, 10, and 20

Estimator	Estimates of true effect: (1) *WLS*	(2) *PET-*	(3) *WAAP*	(4) *WAAP-*	Medians
Part A. Results for k = 5					
(1) Mean	158	358	213	239	226
(2) Median	108	334	132	143	137
(3) Paldam	66	142	80	82	81
(4) *WLS*	76	169	83	99	91
(5) *PET-PEESE*	140	158	141	150	145
(6) *WAAP-WLS*	79	155	77	79	79
Median for true effects	93	164	108	121	
Part B. Results for k = 10					
(1) Mean	133	284	191	199	195
(2) Median	104	267	131	206	168
(3) Paldam	60	131	74	78	76
(4) *WLS*	63	120	80	82	81
(5) *PET-PEESE*	121	176	105	150	135
(6) *WAAP-WLS*	82	141	62	69	76
Median for true effects	93	159	92	116	
Part C. Results for k = 20					
(1) Mean	137	284	182	209	195
(2) Median	99	220	148	174	161
(3) Paldam	60	113	66	79	73
(4) *WLS*	56	120	70	82	76
(5) *PET-PEESE*	105	138	94	117	111
(6) *WAAP-WLS*	73	106	62	65	69
Median for true effects	86	129	82	99	

Note: k denotes the number of empirical studies. Bold highlights lowest MAPE.

Table 6 summarizes the median MAPE and SMAPE for each of the six estimators and across the evolution of research from: 5 studies, 10, and 20 studies. The last column reports the median of the medians. The Paldam rule and *WAAP-WLS* have the lowest median of medians by the MAPE, but this honor goes to *WLS* if SMAPE is used as the measure of predictive accuracy. Because we can never know exactly the 'true' effect, the Paldam rule presents a nice and simple first approximation to the 'truth', one that is hard to beat by any criterion.

Similarly, it seems clear from past simulation studies and these predictive accuracy results that *WLS* and *WAAP* are the best proxies of 'true' effect and that *PET-PEESE* is too

unstable to provide a reliable benchmark (Stanley, 2017; Stanley et al., 2017). One problem with *WAAP* is that cannot be used in many cases where statistical power is low or effects are small. However, using the conditional *WAAP-WLS* estimator, which combines these two approaches, gets around this limitation (Stanley et al., 2017).

Table 6. Summarizing and comparing MAPE and SMAPE medians reported as a percent

		(1)	(2)	(3)	Median of
Number of studies, *k*		5	10	20	Medians
	Part A: Summary of Table 4, median MAPE for each estimator				
(1)	Mean	226	195	195	195
(2)	Median	137	168	161	161
(3)	Paldam	81	76	73	76
(4)	*WLS*	91	81	76	81
(5)	*PET-PEESE*	145	135	111	135
(6)	*WAAP-WLS*	79	76	69	76
	Part B. Parallel results for SMAPE				
(1)	Mean	110	101	99	101
(2)	Median	97	90	89	90
(3)	Paldam	83	75	71	75
(4)	*WLS*	89	67	62	67
(5)	*PET-PEESE*	135	111	99	111
(6)	*WAAP-WLS*	90	73	64	73

Note: the detailed results for SMAPE parallel to Table 5 is presented in Appendix B.

Figure 3: Absolute correlation, alternate estimates of 'true' effect as *k* increases

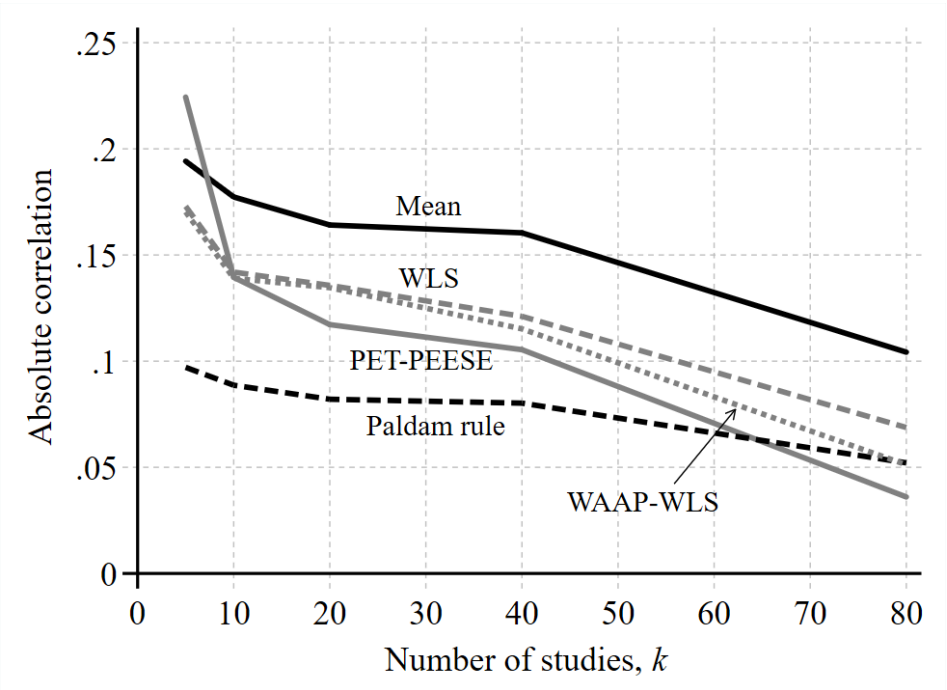


Figure 3 compares the absolute value of the estimated effect for the 50 fields that use correlations as the effect size. The Paldam rule does relatively well because effect sizes are falling, on average. *FAT-PET* performs relatively poorly for small *k*, but then convergences more rapidly as *k* increases. It is important to note that all three meta-averages converge, as they should, but the simple mean remains exaggerated by approximately a factor of two. At the limit, one would hope that all meta-analysis methods approach the 'true' value of the policy parameter of interest, and this seems to be what they do.

6. Research inflation or exaggeration

Ioannidis *et al.* (2017) define research inflation as the degree to which the research record exaggerates the 'true' effect. Following Ioannidis *et al*. (2017), we measure research inflation relative to *WAAP*. In our context, research inflation measures the degree to which the earlier studies exaggerate the estimate of the 'true' effect. In some cases, the early research record will understate the 'true' effect.

We focus on the first 5 studies, which is the most challenging stage given the dearth of data. The degree of research inflation/deflation is presented in Table 7. We find that nearly three-fourths exhibit exaggeration when the simple mean is used to summarize the early research, but this percent goes down, as one would expect, when the Paldam rule or one of meta-analysis estimators is employed. *WLS* does the best in minimizing initial exaggeration as well as initial underassessment. *PET-PEESE* again performs relatively poorly in an environment where the evidence is thin.

Table 7. Research inflation/exaggeration, first 5 studies

All cells in %	(1) Fraction inflated	(2) Fraction deflated	(3) Median inflation	(4) Median deflated	(5) Wrong sign
Unweighted	70	30	242	-66	14
Paldam	52	48	154	-57	14
WLS	66	34	110	-62	13
PET-PEESE	44	56	204	-108	29
WAAP-WLS	61	39	199	-62	14

Table 7 collaborates Ioannidis et al. (2017) that the majority of the initial findings are highly inflated, but this is not true for a notable minority of areas of research. Under- or over-statement of results can emerge from several factors. For example, this development might merely reflect some bias in the early literature that is corrected by the subsequent research record. It may be due to random heterogeneity, or it might reflect genuine heterogeneity whereby the effect gets larger or smaller over time.

Column (4) reports the percentage of areas in which the initial findings get the sign of the effect wrong. It appears that there is about a 14% chance that early literature gets the sign of the effect wrong. Needless to say, this could be a very costly error for evidence-based policy.

7. Discussion and conclusions

"Economists have the least influence on policy where they know the most and are most agreed; they have the most influence on policy where they know the least and disagree most vehemently." —Blinder (1987: 1)

Empirical studies take time to develop and become publicly available. The time delays can be long. For example, on average, it takes about 10 years for a literature to report 10 empirical studies. This is a long time for policy makers to wait, and research consensus will take much longer to emerge. In the meantime, pressing policy decisions have to be made. Is it prudent to wait? Is it the best 'policy' to wait until the research record is settled or at least mature? Fortunately, our survey shows that methods are available that can inform policy even when the evidence base is thin.

From a practical point of view, ease of use and transparency, the Paldam rule offers a simple and surprisingly accurate approach to forecasting what an area of economic research is likely to say in the future. Policy makers can simply scan all of the available, yet small, number of estimates, calculate a simple unweighted average, and then divide by two. This 'half' average can be recalculated as the evidence base deepens. The associated errors seem acceptable, averaging about 67 to 81%. In practice, WLS is virtually as good as or a little better than the Paldam rule, and it has impeccable statistical properties grounded on the authority of the Gauss-Markov Theorem (Stanley and Doucouliagos, 2015; 2017). Thus, a casual reading of economics research would benefit from the routine application of the Paldam rule—just divide by 2. More formal summaries of research for evidence policy should at least calculate

WLS (if not *WAAP*, as well), because there is little to lose by doing so and potentially much to gain. By merely calculating this simple *WLS* weighted average, as much as half of the exaggeration of early research, or more, is typically removed.

Our main conclusion is that there are methods, namely the Paldam rule and the *WLS* meta-average that can be used by policy makers even if the evidence base is thin. This is heartening and suggests that better policy can be made even when the evidence base is insufficiently mature. We also confirm what meta-analysis has long established: simply calculating an unweighted average predictably gives large errors and is not advisable.

The results presented in this survey are also of interest for meta-analysis in general. The median number of studies in our survey is only 20. This suggests that the evidence base for most empirical economics is rather limited. While this may indeed reflect the final size of a research literature, it is entirely possible that when a meta-analysis is conducted, it occurs at a time when the literature is still unfolding. However, meta-analysts often ignore that they are taking a snapshot of the past, which may or may not be indicative of the future. Our survey confirms that effect sizes in most areas in economics are declining. That is, there is a good chance that meta-analyses (or any other review) may be reporting inflated effect sizes. Indeed, many authors of meta-analysis have observed that effect sizes are falling.

As an anonymous reviewer observed, broad meta-analysis surveys such as this one may be viewed as contributions to the history of thought or the philosophy of science. However, instead of using *a priori* logical or normative criteria, as the logical positivists and naïve falsificationists did decades ago, meta-analysis offers an empirical assessment of economics research using the same tools, more or less, as those employed in economics research. We believe that meta-analysis provides a 'naturalistic' philosophical perspective on economic science which is a "turn away from a priori philosophy and towards a philosophical vision that is informed by contemporary scientific practice" (Hands, 2001: 129). "A well-conducted MRA may also serve as the basis for an internal philosophical appraisal of the scientific progress . . . . MRA's potential for deeper philosophical reflection and evaluation has also been largely untapped" (Stanley and Doucouliagos, 2012, 'Coda,' pp. 151-2). For example, logically linked meta-analyses have been used to appraise the natural rate hypothesis and found it wanting (Stanley, 2004; Stanley, 2005; Stanley, 2013).

Along with Ioannidis et al. (2017), this 'meta-science' study identifies low power, publication bias and research exaggeration as severe limitations to the practical policy value of economics research. The current study also identifies a pattern of diminution to this exaggeration as an area of research develops and advances a simple rule and weighted average

to mitigate its potential to stimulate policy miscalculation. Thoughtful surveys of meta-analyses have great potential to detect patterns and limitations in the actual research record and thereby offer practical advice for the advancement of economics research and policy.

"In our view, the central task of meta-regression analysis is to filter out systematic biases, largely due to misspecification and selection, already contained in economics research" (Stanley and Doucouliagos, 2012, p.12). But there are limits to what any research tool can do, and it is unlikely that the research base will be sufficiently informative as to allow meta-analysis to identify and remove all biases reliably. The findings in this paper suggests that limitations of early research may be more serious than previously recognized. Our survey highlights the need for reviewers and meta-analysts to focus more attention on parameter stability.

Acknowledgements:

Zohidjon Askarov provided assistance with coding and analysis. This project was funded by the Deakin Lab for Meta-Analysis of Research (DeLMAR). The paper has been presented at the Conference on the Political Economy of Public Policy 2017, Ariel University (Israel), and at the MAER-Net Colloqium 2017 at Zeppelin University (Germany). We are grateful to the comments made by the participants.

References

Bailey, C.D., Hasselback, J.R., Karcher, J.N. 2001. Research misconduct in accounting literature: a survey of the most prolific researchers' actions and beliefs. *Abacus* 37 (1), 26–54.

Blinder, A.S. 1987. *Hard heads, soft hearts: Tough-minded economics for a just society*. Addison-Wesley: Reading, Massachusetts.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T-H., Huber, J. et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 1433.

Cohen, J. 1988. *Statistical Power Analysis in the Behavioral Sciences*, 2nd edn, Hillsdale, NJ: Erlbaum.

Dewald, W.G., Thursby, J.G., Anderson, R.G. 1986. Replication in empirical economics. The Journal of Money, Credit and Banking Project, *American Economic Review* 76, 587-603.

Doucouliagos, H., Paldam, M. 2015. Finally a breakthrough? The recent rise in the size of the estimates of aid effectiveness. *Handbook on the economics of foreign aid*, Edward Elgar, Cheltenham, Eng., pp. 325-349.

Evanschitzky, H., Armstrong, J.S. 2010. Replications of forecasting research. *International Journal of Forecasting* 26, 4–8.

Hands, D.W. 2001. *Refection without Rules*, Cambridge: Cambridge University Press.

Heckman, J. 2000. Microdata, Heterogeneity and The Evaluation of Public Policy. Bank of Sweden Nobel Memorial Lecture in Economic Sciences December 8, 2000 Stockholm, Sweden.

Hubbard, R., Vetter, D.E. 1992. The publication incidence of replications and critical commentary in economics. *The American Economist* 36, 29–34.

Ioannidis, J.P.A., Stanley, T.D., Doucouliagos, H(C). 2017. The power of bias in economics research. *The Economic Journal* 127, F236–F265.

Ioannidis, J.P.A., Trikalinos, T.A. 2005. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomised trials. *Journal of Clinical Epidemiology* 58, 543–549.

Kim, S., Kim, H. 2016. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting* 32, 669-679.

Kitchenham, BA, Pickard, L., MacDonell S. 2001. What accuracy statistics really measure. *IEE Proceedings—Software* 148 (3), 81–85.

Kolassa, S., Martin, R., 2011. Percentage errors can ruin your day (and rolling the dice show how. *Foresight: The International Journal of Applied Forecasting* 23 (Fall), 21–27.

Makridakis, S., Wheelwright, S.C., McGee, V.E. 1978. *Forecasting: methods and applications*, Wiley: New York.

Miguel, E., Christensen, G. 2017. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* 56, 920-80

Moreno, S. G., A. J Sutton, A E Ades, T. D. Stanley, K. R Abrams, J. L Peters, N. J Cooper. 2009. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* 9, 2 http://www.biomedcentral.com/1471-2288/9/2

Office of Management and Budget 2012. Memorandum to the heads of executive departments and agencies, Jeffrey D. Zient, Washington, D.C. 20503

Paldam, M. 2017. A model of the rational economist, as researcher and policy advisor. *European Journal of Political Economy*, 54, 6-15

Rosenberger, R.S., Stanley, T.D., 2006. Measurement, generalization and publication: sources of error in benefit transfers and their management. *Ecological Economics* 60, 372–378.

Stanley, T.D. 2004. Does unemployment hysteresis falsify the natural rate hypothesis? A meta-regression analysis. *Journal of Economic Surveys*, 18(4): 1-28.

Stanley, T.D. 2005. Integrating the empirical tests of the natural rate hypothesis: A meta-regression analysis. *Kyklos*, 58(4): 587-610.

Stanley, T. D. 2013. Does economics add up? An introduction to meta-regression analysis. *European Journal of Economics and Economic Policy* 10: 207-220.

Stanley, T.D. 2017. Limitations of PET-PEESE and other meta-analysis methods. *Social Psychology and Personality Science*, *8*(5), 581-591.

Stanley, T.D., Doucouliagos, H(C). 2012. *Meta-Regression Analysis in Economics and Business*, Routledge, Oxford

Stanley, T.D., Doucouliagos, H(C). 2014. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods* 5, 60-78.

Stanley, T.D., Doucouliagos, H(C). 2015. Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine* 34, 2116-2127.

Stanley T.D., Doucouliagos, H(C). 2017. Neither fixed nor random: Weighted least squares meta-regression analysis. *Research Synthesis Methods* 8, 19-42.

Stanley, T.D., Doucouliagos, H(C)., Ioannidis, J. 2017. Finding the power to reduce publication bias. *Statistics in Medicine*. DOI: 10.1002/sim.7228.

Tofallis, C. 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society* 66 (8), 1352-1362.

UK Cabinet Office. 1999. Modernising government, Cm 4310, London: The Stationery Office, www.archive.official-documents.co.uk/document/cm43/4310/4310.htm .

Viscusi, W.K. 2017. Best estimate selection bias of estimates of the value of a statistical life, *Journal of Benefit-Cost Analysis*. Forthcoming.

Viscusi, W.K., Aldy, J.E. 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27 (1), 5–76.

Young, N.S., Ioannidis, J.P.A. and Al-Ubaydli, O. Why current publication practices may distort science. *PLoS Medicine, 5*:1418–22, http://dx.doi.org/10.1371/journal.pmed.0050201.