Aarhus, 28/3-2012

An essay about publication bias and strikes

Martin Paldam

Department of Economics and Business, Aarhus University,
Bartholins Allé 10, Building 1322, 8000 Aarhus C, Denmark.
E-mail: mpaldam@econ.au.dk, URL: http://www.martin.paldam.dk

Abstract:
This essay revisits a paper published by Peder J. Pedersen and myself three decades ago. It contains an empirical cross-country study of the macroeconomics of strikes, and a story about the literature, which we found intriguing. It dealt with a problem in the political economy of economic research. Now we know that the problem was publication bias. The story is one aspect of a common finding in the meta-analysis literature that has flourished the last five years.

1. The primary study and a story of the literature

This essay discusses a paper Peder J. Pedersen and I published in 1982 no less than 30 years ago. P&P contains a primary empirical study of the macroeconomics of strikes on a data panel of 17 countries over 30-35 years. In addition P&P tells a naughty story about the literature. The story points to a problem in our science. It is downplayed, but we giggled when we wrote certain paragraphs. At that time no formal technique existed to study the problem. In the meantime the technique has been developed as discussed in sections 2 and 3.

1.1 Our primary study of the data: The macroeconomics of strikes

P&P used econometrics of the 1981 vintage, which did not include panel regressions. What we did was to run independent OLS-regressions using precisely the same model on the data for each country, and then we studied the cross-country pattern by casual methods.

The results reached showed a main cross-country pattern and one outlier. The main pattern is not found in the data of the outlier, and there is no trace of the outlier pattern in the data for the other countries:

Main pattern. Strikes moves in waves which might be started by a wage or a strike push. Strikes cause nominal wage increases, so strike-waves are parts of joint wage-strike waves. Thus, the main story is about nominal *wage-competition.* It is weakly reflected on the real level too, so strikes are positively related to real wage rises in most countries.

Outlier. In one country strikes are negatively related to real wage rises. Strikes occur when real wages increase less than they used to. The outlier can be explained in two ways: It may be due to (i) a *random quirk* in the data, or it may be genuine and due to (ii) a *special institution* that makes the labor market of the outlier country react differently.

When reading the paper I was visited by one of these thoughts that really should perish. I kept thinking about what would have happened if we had had access to the sparkling new tools of the 2011 vintage. We would surely have got our result a few days earlier, and we could have presented them in a much more impressive way, but my problem was that I could not get rid of the feeling that we would have reached the same conclusions.

1.2 The story about the literature

In addition to our own study we also surveyed the literature. This led to the story that will be the theme of the rest of the essay. At that time the macroeconomics of strikes was a field with

a nice little literature. It was like the story of Snow White and the seven dwarfs. One paper is tall and beautiful and the other seven are more modest in size and beauty. The Snow White paper (Ashenfelter and Johnson 1969) came first. A&J looked as a paper should: It was written by young stars using US-data and published in the American Economic Review.[1] It starts with a formal theory of *grumpiness*. When workers do not get the wage increases they expect, they turn grumpy and strike. This theory appears reasonable, and when it is presented with an adequate number of Greek letters, the profession is happy to believe. The theory is operationalized and tested using 1968 vintage econometrics.

However, there is a problem – it dealt with the outlier. A&J do not try to explain why the US is different. No country is explicitly mentioned in the paper. It is taken for granted that it the theory is tested on a fine data set. When rereading the theory nothing singles it out as a US-theory. It appears likely that when the data for 17 countries are examined one outlier should occur. So perhaps the quirk theory applies? Be it as it may, this star-paper imbued the researchers in the field with an A&J-prior for *the right sign*.

Now to the interesting part of the story: The next seven papers managed to find the right signs! They were written by perfectly decent researchers dealing with other countries or time periods. None of the papers used exactly the same model as A&J, but somehow it looked like almost the same, and they (largely) confirmed and strengthened the A&J prior[2] – even when it was the outlier: Clearly something odd was going on.

Over the years we have all heard remarks – often as an aside – about the element of make believe in empirical macro research. We all know that coefficients are less significant than advertised. The steadily growing kit of econometric tools can be used for truth finding, or, as suggested by Ronald Coase, to 'torture the data long enough' so that 'it will confess'.

The story thus calls for some comments – a small one and five large: The small one is that our paper was not written as a critique of any of these authors.[3] The five large comments can be expressed as questions: (Q1) How easy is it to find the right signs? (Q2) Is the story part of a broader problem? (Q3) Does a technique exist to reveal when this problem occurs? (Q4) What would have happened if economic interests had been involved? (Q5) The old tools issue mentioned. Before I return to these questions in section 4, a detour will be made to the modern literature on meta-analysis.

1. In Harzing's Publish or Perish Ashenfelter and Johnson (1969) have acquired 555 citations, while Paldam and Pedersen (1982) have reached 60 citations.
2. It is possible that the story has a simple institutional explanation. Perhaps A or J refereed most of the papers.
3. We re-found the data regularity reported by A&J in the US-data. None of the seven authors tried to hide their theoretical prior. They looked for the right signs, and their quest was rewarded.

2. Meta-analysis: Analyzing the β-literature with *N* estimates of β

Meta-analysis was developed in medicine, but the method has now been adopted for the use in economics. Here it is common that a set of L papers brings estimates of a parameter, β, which is defined in the same way, implying that it is the *same* β.[4] The L papers of the β-literature contain N estimates.[5] The meta-technique is used to extract the *best* average of all these estimates. The present meta technique was only developed in the last 5 years.[6]

2.1 The framework: A quantitative survey of the β-literature

A new primary empirical study gives a contribution to the β-literature of about $1/L$. If L is large the contribution is small. It is possible – perhaps even likely – that it would be a much bigger contribution to extract the 'best' average value of β from the L studies.

This is precisely what the meta-analysis does by compiling a data set covering the β-literature. In addition the data set allows the analyst can ask questions to the literature such as: Do structural shifts due to theoretical and econometric progress occur? Are results published in top journals different? Do identifiable schools produce different results? Do researches sponsored by a β-related industry find results with a bias that the industry likes? Are results from different countries and time periods different?

The L papers give N estimates, $b_1, \ldots, b_N$ of β, which have to be converted to the same scale, using e.g., partial correlations. Each estimate comes with a standard error, s_i, so that the t-ratio $t = s/b$, and the precision $p = 1/s$ of the estimate can be calculated. Thus the coding covers the B, S, T and P vectors with N elements each. In addition the (N x M) matrix Ω, is coded. Here each row gives a list of characteristics of the estimates, using the same check list. If N is large, such as 500, the check list typically gets rather long, such as $M = 100$.

Most of the t-ratios in the T-set are typically well above 2. This suggests that β is known with great precision. Unfortunately, this is too good to be true!

Meta-analysis has two levels. ***The basic level*** consists of four steps: (s1) a search for the full β-literature; (s2) a coding of the literature; (s3) a study of the distribution (funnel) of

4. Like A&J the typical papers in the β-literature present a *general* model that contains the 'deep' parameter β, then a data set is chosen, and the author goes ahead estimating β from these data. It is rarely asked if the β reached is special for that data set only, so it is implied that the paper tries to reach a general value for β.

5. N is typically 10 times larger than L, so there is typically a 'cluster' of about 10 estimates in each paper. This explains why clustering-corrected standard errors are commonly used in meta-analysis.

6. The first paper proposing that economists use meta-analysis was Stanley and Jarrell (1989). The start was slow, only a handful of papers were made in the 1990s, but now at least 300 meta-studies in economics have appeared. In a few months the first textbook on meta-analysis in economics will be on the market (Stanley and Doucouliagos 2012).

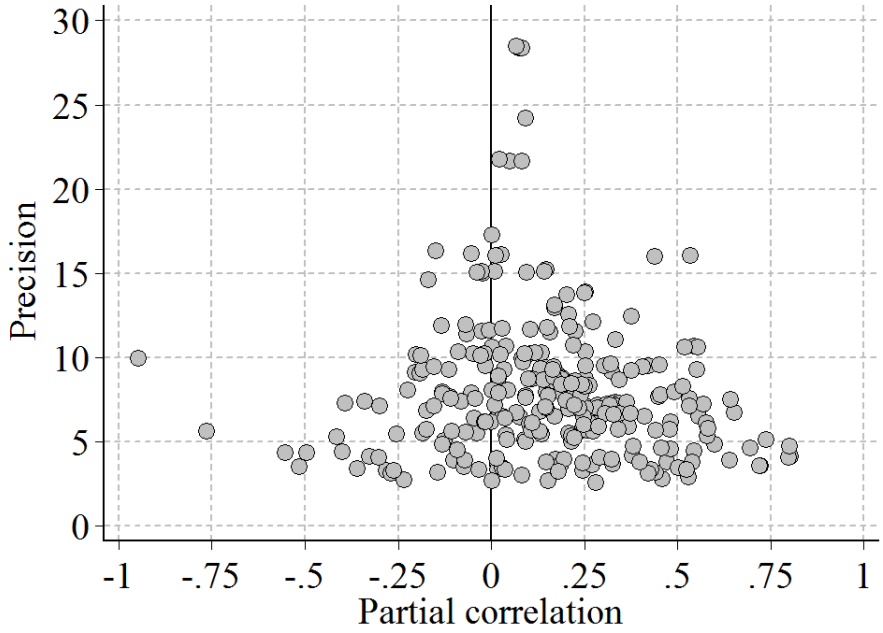
the estimates; and (s4) the estimate of the meta-average. ***The secondary level*** studies the variation in the results. This essay only discusses the basic level.

2.2 Steps 1, 2 and 3: Literature, coding and the funnel

It is a large chunk of work to collect and (especially) to code the β-literature. Once it is done, it can be published and checked by others. It is difficult to get everything right, and one can go on chasing coding errors for a long time. A careful coding will ensure that the fraction of errors is well below 1% of the cells. If the errors are stochastic, the marginal benefits of further error chasing soon become insignificant.

The next step is to study the distribution of the results. It is done in the form of the funnel-diagram that is the (*b*, *p*)-scatter. It gives the estimated *b*s over their precision *p*, but it is turned 90 degrees to be easy to read. It looks as a funnel because the more precise estimates are closer together. Given the high level of t-ratios funnels should be lean. The average result should be the same at all levels of precision, so funnels should be symmetric. As the reader may have guessed empirical funnels are neither lean nor symmetrical, but have the two properties listed in Table 1.

Figure 1. An example of a typical funnel



Note: The figure shows the first 245 reported estimates of aid effectiveness on growth recalculated to partial correlations (Doucouliagos and Paldam 2011). The funnel 'explains' the estimates by their precision. The average should be the same at each level of precision, so the axis of symmetry should be vertical. It is easy to see why the FAT rejects symmetry: Compared with the top there are too many positive estimates at the bottom.

Table 1. Two typical properties of empirical funnels

	Property	Funnel looks	Explanation	Test
(A)	Excess width	Too wide at base	Estimates differ more than suggested by t-ratios	MST
(B)	Asymmetry	Has thin parts	The funnel base is skew relative to the top	FAT

Figure 1 is an empirical funnel. It is chosen for being a typical specimen with the two properties. The estimated b's are converted to the same scale using partial correlations. The theoretical range of the partial correlation is from −1 to +1. The figure covers estimates from −0.95 to +0.82. They pertain to be estimates of the same parameter, and most are published in perfectly decent journals. As the reader can calculate about 60% of the estimates are significant.[7] However, the most precise studies find rather small estimates.

For more examples – including more extreme ones – showing how funnels look the reader should consult Stanley and Doucouliagos (2010). One of the reasons the profession should look at meta-studies is the amazing width of funnels. Funnels help building some of the skepticism necessary for a researcher. The excess width of funnels points to a fact we often manage to forget: It is far too easy to produce nice t-ratios especially in small samples.

Most literatures studied have asymmetric funnels. This is a finding *which should be explained*. Section 3 looks at the main explanation.

2.2 The FAT-PET meta regression analysis

The final step of the basic meta-analysis is to calculate the PET estimate of the meta-average, b_M. It is done jointly with the FAT by the FAT-PET estimate (1) from Stanley (2008). The funnel is asymmetric if $\beta_{FAT} \neq 0$, and then the PET adjusts the average for the asymmetry.[8]

(1) $b_i = b_M + \beta_{FAT}\, s_i + u_i$, estimated by WLS with precisions weights.

Or $b_i = b_M + \beta_{FAT}\, /p_i + u_i \;\; \rightarrow b_M$ for $p_i \rightarrow \infty$

The intuition behind estimate (1) is that even if the funnel is asymmetric ($\beta_{FAT} \neq 0$) the relation converges towards $b_M \approx \beta$ as precision rises. In the case shown, the convergence is towards 0.08.[9] Equation (1) is related to the Heckman correction for selectivity bias, see

7. The lines of significance (for $t = 2$) are the two curves $b = \pm 2/p$ that are symmetrical hyperboles around the p-axis, looking like a 'lazy tent.' About 60 % of the observations are outside the 'tent' and thus significant.
8. A number of simulation studies have showed that if the asymmetry is due to publication bias the correction is very good. That is, the PET gets much closer to the true value of β than averages without the FAT correction, see Stanley (2008) and Callot and Paldam (2011).
9. For the problem covered by Figure 1, we have collected N = 1,361 estimates. When N grows b_M falls, but from about N = 400 b_M stabilizes at 0.03.

Stanley and Doucouliagos (2011). The Heckman procedure is better if the selection process is known, but it does not work when the process is unknown – here (1) can still be used.

Once the literature is coded there is one and only one estimate of the FAT-PET. It completes the basic level of the meta-analysis. It is almost objective because everything is controllable. If two meta-studies of the same β-literature are done independently, they researchers will find almost the same literature, and do almost the same coding, and thus estimate virtually the same FAT-PET. If they check and correct the mistakes of each other, they should reach *exactly* the same FAT-PET.

Doucouliagos and Paldam (2011) – the source of Figure 1 – documents a case a meta-study of 68 papers where three recodings were made with virtually no effects on the estimated meta-average.[10] This appears to be the typical outcome of recoding. The basic level of meta-analysis is by now a well-established procedure giving very robust results.

3. Funnel asymmetries and publication bias

Empirical macroeconomics has the iceberg property. The regressions published are a small fraction, 1/*J*, of the regressions done, where *J* is the mining ratio. It is probably as high as 25. Economic theory predicts that it is rising: The marginal benefit of regression *J* is probably fairly constant and the marginal costs are rapidly falling.[11]

One reaction to the alarming rise in *J* is to increase the costs of running regressions by demanding that regressions are fancier. This helps explaining the rapid increase in the supply to the market of such estimators. However, the competition between econometric packages causes new estimators to be quickly included. Dozens of estimators not known three decades ago are now just a few keystrokes away.

3.1 Is the visible top of the iceberg representative of the whole berg?

Once this question is asked, the answer is obvious: Not very likely!

10. D&P published a meta-study in 2008, which another researcher wanted to disprove, so he organized a very critical re-coding finding about 0.2 % coding errors. They had virtually no effect on the FAT-PET estimate. Then D&P employed a new team of research assistants to do a third recoding, once again with no effects.

11. See Paldam (2012) for a theory of the mining ration. Since P&P the cost of running a regression has fallen at least 50 times (or by 12 – 14 % per year). Back 30 years ago data had to be found in tables published in books, and they had to be punched in and stored on magnetic tape, and it took over 5 minutes to run a regression on a mainframe – mostly due to input and output constraints. You had only access to the mainframe computer a few times a day. Output was on big sheets of special paper, with green lines, which researches kept in big piles in their office. Now the data are downloaded to your pc, and by a few keystrokes the next regression flashes across the screen.

The difference between the average result *reached* and the average result *published* is the *publication bias*. Below a measure of the bias will be given. Researchers react to incentives, and they have no incentive to choose representative estimates. They will surely choose estimates that are *better* than the average one. Consequently, publication bias must be common. They are caused by priors for certain properties of the results, leading to choice rules. The two most likely rules give the most common publication bias: Polishing and censoring, defined in Table 2.

Table 2. The two main types of publication bias

Name	Prior for	Choice rule	Effect on funnel
Polishing	Clarity	Size of t-ratio	Extra width
Censoring	Right results	Size of *b*	Asymmetry

Polishing means that the estimates are chosen by their t-ratio. It follows that published t-ratios are too high. Also, the mining ratio is relevant for calculating the degrees of freedom, *Df*s, of the tests. Imagine that $J = 25$ and each regression estimates 10 quantities, so that 10 *Df*s are used. This means that the loss of *Df*'s is $25 \times 10 = 250$. If you have managed to collect 100 observations in your data set $Df = 100 - 250 = -150$. It is not clear (to me) how this should be understood, but it must surely mean that the t-test done are less convincing than advertised. With *J* rising there is a drift into the land of make believe.

Censoring means that the published *b*s are chosen for their size – it introduces a loop from the desired coefficient to the published result. This is deeply problematic, but unavoidable. If ethical rules against such loops are announced, it gives a strong moral hazard problem in a world where researchers must publish. Thus, it is better to take it for granted that these things happen, and try to correct for the problem.

3.2 A simple example of a prior generating a large publication error

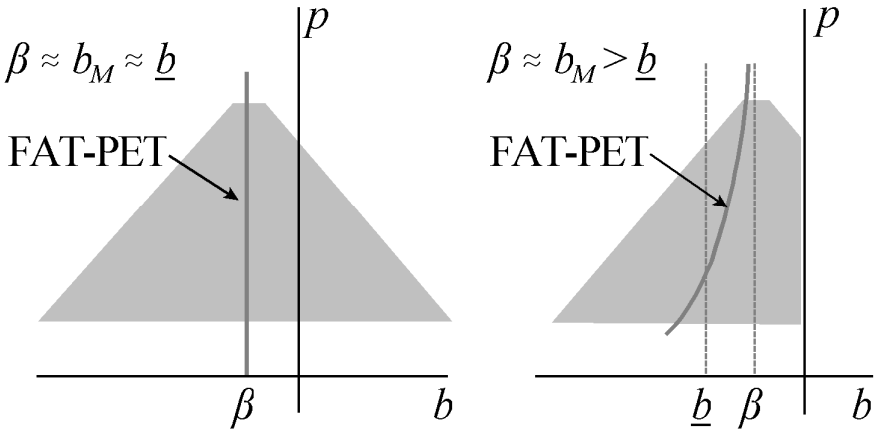
Figure 2 shows a stylized version of the classical example (from microeconomics) of the price elasticity for beer.[12] Most researchers in the field have a degree in economics. We preachers of economic theory have taught our students that such elasticity must be negative, and common sense tells us that it not numerically large. So we have provided the researchers with a sign prior – a prior that applies even for estimates on small data samples.

12. Though I have seen the funnel, the paper has not yet been completed.

To estimate the elasticity requires an identifying assumption. A range of possibilities exists, and it is difficult to choose the right one. Also, the form of the estimating equation has to be chosen, and the relation has to be controlled for alternative beverages, the weather, the fraction of Muslims, the number of sport events requiring six-packs, etc. It gives a range of estimates, as shown on Figure 2. Some of these estimates are positive.

Figure 2. An illustration of the censoring story: A strong prior for negative results

a. Funnel of all estimates b. Censoring by a negative sign prior



Note: β the true value, $\underline{b}$ the arithmetic average and b_M the meta-average from equation (1).

If all estimates made were available – as depicted on Figure 2a – the funnel would be symmetrical. The FAT-PET of equation (1) would find that $\beta_{FAT} \approx 0$, so it would be a vertical line as shown. The arithmetic average, $\underline{b}$, and the meta-average, b_M, would be the same, and they would both be fine estimates of the true value β.

However, due to the sign prior the researchers will not publish a positive elasticity. If they did their colleagues and their old professors will laugh. So the published results come to look as Figure 2b. The censored funnel has the arithmetic average $\underline{b} < \beta$. The FAT-PET now estimates a significantly negative β_{FAT} indicating an asymmetry, but the PET will still converge to the true value β, as shown. Consequently, the meta-study will find $b_M \approx \beta$.

In a traditional survey the possibility of publication bias is disregarded, and it typically ends with some common average, such as $\underline{b}$. This is an exaggerated value of β. Hence, a measure, *PB*, of the publication bias can be estimated as:

(2) $PB = \underline{b}/b_M \approx \underline{b}/\beta$

The illustration on Figure 2b finds a publication bias of two. Douliagos and Paldam (2011) report estimates of $\underline{b} \approx 0.16$ and $b_M \approx 0.08$ for the data presented on Figure 1. Thus it gives a *PB* value of two as well. Doucouliagos and Stanley (2012) have collected estimates of *PB*s from meta-studies of 87 literatures. The distribution of the publication bias in different literatures is quite wide, but two appears a fairly common value.

Also, PB ≈ 2 corresponds to a folk theorem in the meta-research community: If you consider the average paper in a literature, you do not know, expect a publication bias of two.

4. Conclusion

All sciences demand that results are independently replicated, i.e. by other researchers on new data, before they can be trusted. In macroeconomics data-mining is fact of life, so both the published coefficients and their t-ratios are exaggerated. It means that independent replication is particularly important.

4.1 Back to the P&P story: Seven independent replications that failed

When we (P&P) surveyed the macro strike literature three decades ago, we were puzzled. It looked like a brilliant study reaching results that was confirmed by no less than seven independent replications! So, if we had not made our own primary study we would have had to conclude that the A&J model did a good job explaining the macro-pattern in strikes.

However, our primary study showed something very different. Consequently, the story of the literature also told a different story. Somehow the process of independent replication came to work in the reverse. It is story of a high status paper that had the bad luck to deal with an outlier, but nevertheless managed to establish a sign prior. The seven new researchers did not try to test the A&J-result on new data. They tried to confirm the A&J sign prior – and they succeeded!

Seven papers are not enough for a meta-study, but if 25 papers had been published, all suffering from the A&J-prior, they would have generated a nicely asymmetric funnel. And we would have been able to tell, just from the literature study, what was going on.

4.2 Section 1 posed five questions – they can now be answered:

(Q1) How easy is it to find the right signs? Answer: It is far too easy. I have seen about 50 meta-studies in macroeconomics and most of these had a clear censoring asymmetry. Most

censoring is around zero, so it appears that sign priors are common.

(Q2) Is the story part of a broader problem? Answer: It is a part of the publication bias problem. It is large in economics – notably macroeconomics.

(Q3) Does a technique exist to reveal when this problem occurs? Answer: Yes, the meta-technique is now available, and the FAT is a powerful test. Once an asymmetry is detected it has to be interpreted, but if the asymmetry is a (full or partial) censoring at zero and there are reasons to expect such a censoring, a likely explanation has been found.

(Q4) What would have happened if economic interests had been involved? Answer: Many cases are reported where a group of researchers has an interest in certain results. It is normally quite easy to predict what bias it will give,[13] and a number of formal tests are available. They normally confirm the expected bias, but it is often moderate in size.

Finally, there was the 'old tools' question: (Q5) If we (P&P) had had access to the super techniques of today, would we have produced different results? Answer: One way to study that – in a broader setting – is to ask the question at the secondary level of the meta-analysis: How much of the variation between the results in the typical β-literature is due to statistical technique? The studies I have seen all find that it is a small part.

13. The simplest devise is to color the *b*s from the said researchers in a different color and then look at the funnel. This is often enough, to see that there is a problem.

Old references: A&J, P&P and the seven dwarfs[14]

Ashenfelter O., Johnson, G.E., 1969. ***A&J***. Bargaining Theory, Trade Unions, and Industrial Strike Activity. *American Economic Review* 59, 35- 49

Bentley P., Hughes, B., 1970. Cyclical Influence on Strike Activity: The Australian Record 1952-68. *Australian Economic Papers* 9, 149- 70

Edwards, P. K., 1978. Time Series Regression Models of strike Activity: A Reconsideration with American Data. *British Journal of Industrial Relations* 16, 320-34

Hibbs, D.A., 1976. Industrial Conflict in Advanced Industrial Societies. *American Political Science Review* 70, 1033-58

Paldam, M., Pedersen, P.J., 1982 ***P&P.*** The Macroeconomic Strike Model: A Study of Seventeen Countries, 1948-75. *Industrial & Labor Relations Review* 35, 504-21

Pencavel, J.H., 1970. An Investigation into Industrial Strike Activity in Britain. *Economics* 37, 239-56

Snyder, D., 1975. Institutional Setting and Industrial Conflict: Comparative Analyses of France, Italy and the United States. *American Sociological Review* 40, 259-78

Vanderkamp, J., 1970. Economic Activity and Strikes in Canada. *Industrial Relations* 9, 215- 30

Walsh, W.D., 1975. Economic Conditions and Strike Activity in Canada. *Industrial Relations* 14, 45- 54

New references:

Callot, L., Paldam, M., 2011. Natural funnel asymmetries. A simulation analysis of the three basic tools of meta-analysis. *Research Synthesis Methods* 2, 84–102

Doucouliagos, H., Paldam, M., 2011. The robust result in meta-analysis of aid effectiveness: A response to Mekasha and Tarp. *Economics Working Paper* 2011-15

Doucouliagos, H., Stanley, T.D., 2012. Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys* doi: 10.1111/j.1467-6419.2011.00706.x

Paldam, M., 2012. The mining ratio in applied macroeconomics. March p.t. working paper

Stanley, T.D., 2008. Meta-regression methods for detecting and estimating empirical effect in the presence of publication bias. *Oxford Bulletin of Economics and Statistics* 70, 103-27

Stanley, T.D., Doucouliagos, H., 2010. Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys* 24, 170–91

Stanley, T.D., Doucouliagos, H., 2011. Meta-Regression Approximations to Reduce Publication Selection Bias. Deakin University, Melbourne, Australia. Economic series, SWP 2011/4

Stanley, T.D., Doucouliagos, H., 2012. *Meta-Regression Analysis in Economics and Business*. Routledge, (announced to appear 13th July)

Stanley, T.D., Jarrell, S.B., 1989. Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys* 3, 54-67

14. P&P surveys seven more papers that consider other aspects of strikes.