The Power of Bias in Economics

That power which erring men call Chance.
– John Milton (1631, *L'Allegro*).

Tom Stanley
Julia Mobley Professor of Economics,
Hendrix College, USA:
Stanley@Hendrix.edu

with Chris Doucouliagos
and John P.A. Ioannidis

This project would not have been possible without with you!

- We are very grateful for the unbelievable cooperation from **you**.
- **159** meta-analyses datasets.
- **Thank you Very much for your help!**

Origins of this Project

- John Ioannidis’s Keynote at last year’s MAER-Net Colloquium in Athens. . . .
  - John’s ‘Why most published (medical) research findings are false’
    - **Ans: lack of statistical power and potential biases**
  - His ‘How to make more published research true’
    - **Ans: radical revision of our research incentive structure and to promotion and tenure decisions**
- Follow-up MAER-Net online discussion

So what did we actually do?

- Calculate both the **median power** and the proportion of research results that are **adequately powered** (using Cohen’s convention $\geq$ 80%)
- Make conservative assessments of the magnitude of ‘**research inflation’** across these 159 fields.
  - Possible sources of this **exaggeration of effects**: publication selection, small-sample, selective reporting, confirmation and/or experimenter bias.
- Offer a new approach to accommodating these reporting/selection biases—**WAAP**, weighted average of the adequately powered.

Statistical Power

- Power is $1-\beta$; where β is the probability of a type II error. The type II error is the mistake of accepting that there is no effect when, in fact, there is.
- Power is the probability that we can detect what we seek.
- It is analogous to the power of a telescope.

Why is Power Important?

“Unless (we) begin to incorporate methods for increasing the power of (our) studies, the **published literature is likely to contain a mixture of apparent results buzzing with confusion.** . . .Not only do **underpowered studies lead to a confusing literature** but they also create a literature **that contains biased estimates** of effect sizes” (Maxwell, 2004, p.161).

- Without power, a single empirical finding is:
  - essentially worthless
  - as likely bias than informative.
- It is **power, not p-value**, that is the real metric of the importance of an empirical result.
- Unlike low p-values, high power cannot be the result of the selection of biased effects . . . . .

How can power be calculated?

- Hypothetically, for a given difference of interest.
  - Say, for example, you want to detect the difference between a zero price effect and a small one—*i.e.*, elast. = {0 vs -0.2}, and you know the SE or the n.
  - From the '**true**' effect. . . . But how to estimate it?
  - **Primary studies** cannot know the 'true' effect; that's what they are seeking. Ex post, primary studies can only calculate power in a circular, meaningless way.
  - **Meta-averages** are better because they use much more information and average across many studies, their sampling errors and their biases.

How do we calculate power?

Ex post, from over 6,700 studies aggregated into 159 meta-analysis and for more than 64,000 estimates.

To be conservative, we use ‘fixed-effect’ MA or our new **unrestricted WLS—**Stanley and Doucouliagos (2015)

- Not random-effects or the simple average: both are much more biased if there is publication bias (**PB**).
- Fixed-effect (**WLS-FE**) is also biased with PB, but less so; thus will over-estimate the power of economic estimates.

For robustness, we use 3 other proxies for true effect:

- **Top 10%:** WLS-FE of the most precise 10%
- **Top 1:** the single most precise estimate
- **PET-PEESE:** PB corrected estimate, Stan & Doc (2014)

How do we **measure** power?

In two ways, by:

- The proportion that have adequate power: relative to Cohen’s widely accepted standard $\geq$ 80%.
- The median power for each area of research.

What did we find?

- The typical area of economics has about 90% of its estimates underpowered.
  - That is, the median proportion that is adequately powered is, at best, just over 10%--10.5% (**WLS-FE**); 6.5% (**Top 10%**); 5.8% (**PET-PEESE**)
- Median power is, at best, 18%.
  - That is, the median of 159 median powers is: 17.9% (**WLS-FE**); 11.1% (**Top 10%**); 10.7% (**PET-PEESE**); 8.1% (**Top 1**).

Figure 1: Percent of Empirical Estimates Adequately Powered

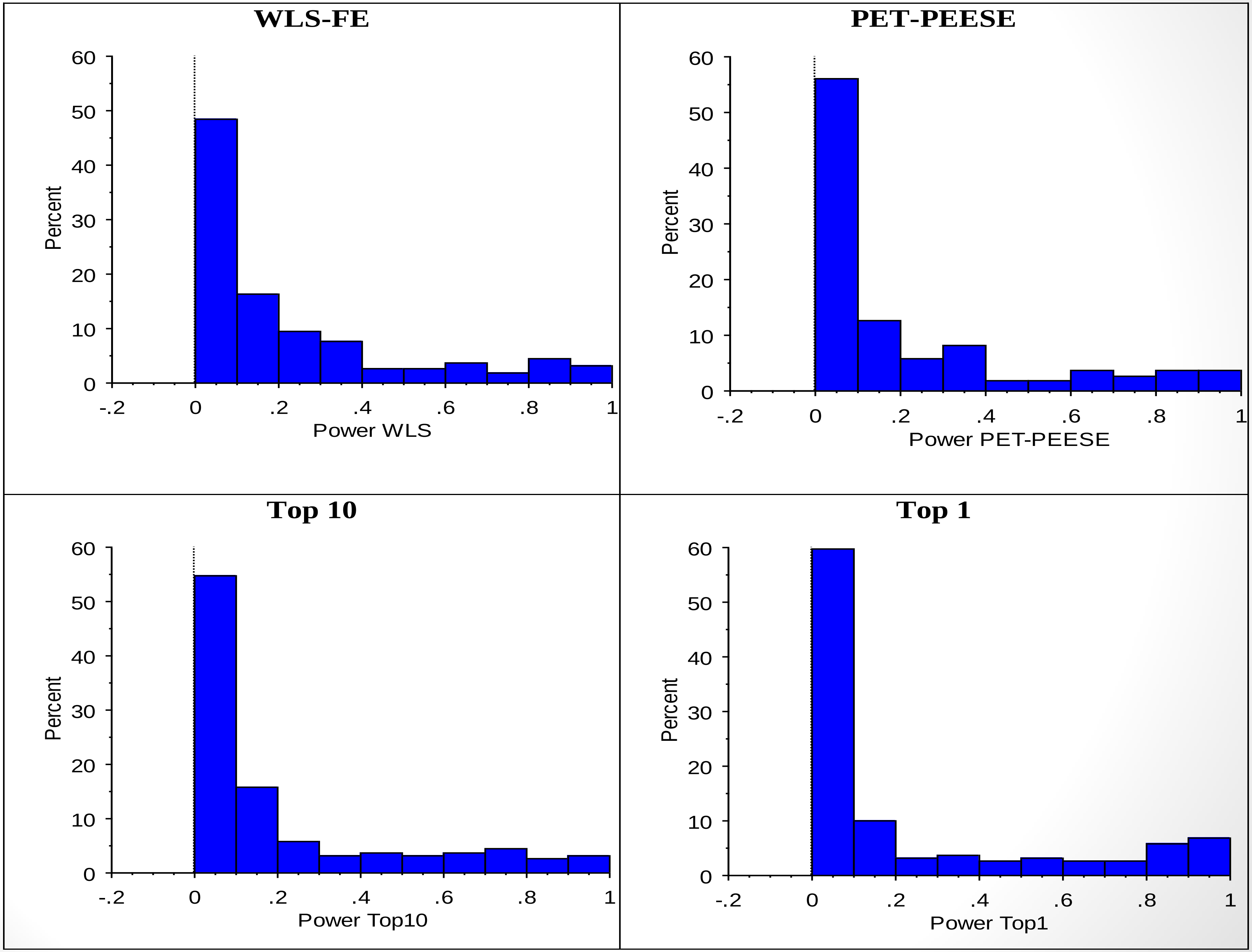
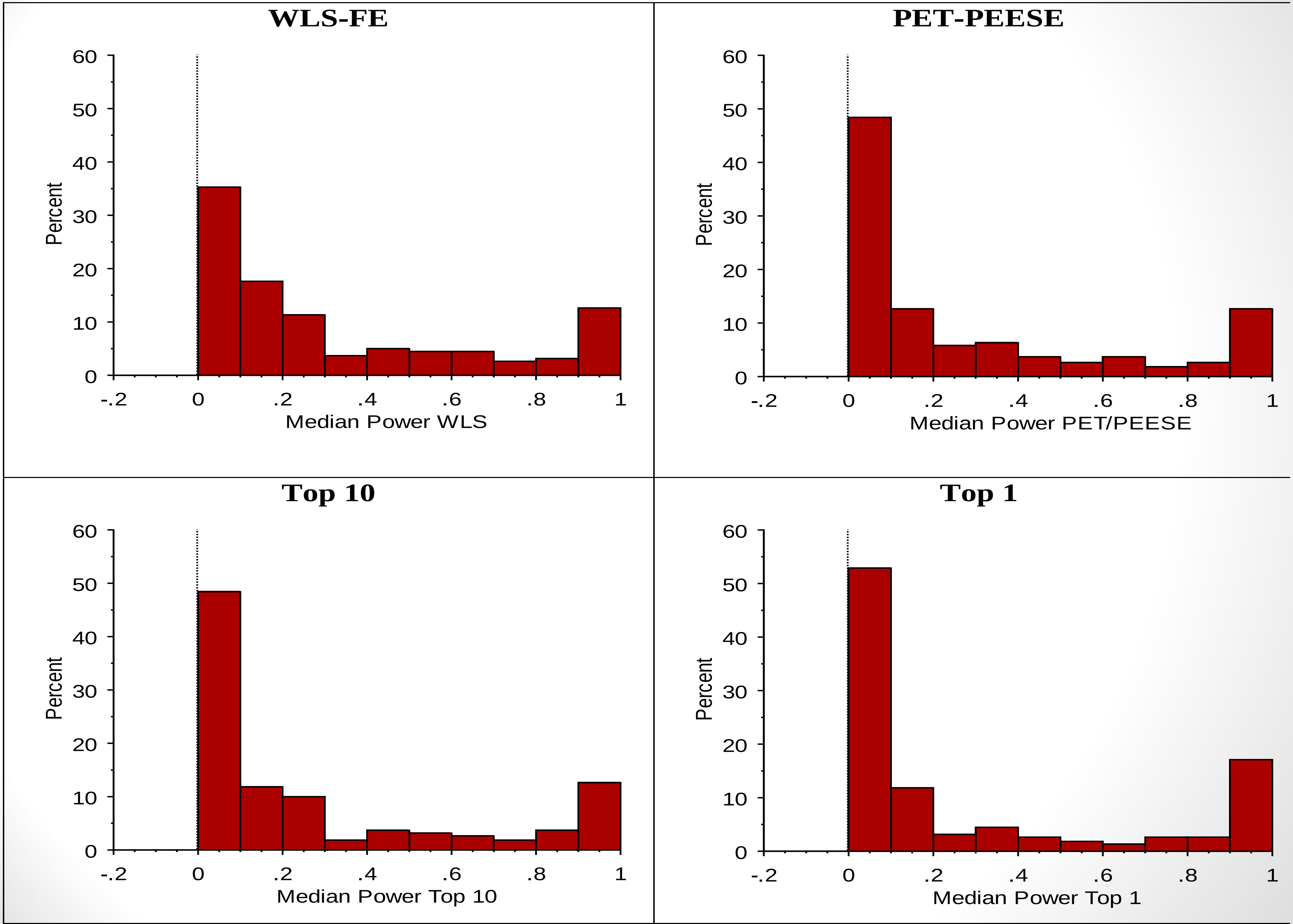


Figure 2: Median Power across 159 Meta-Analyses



Impotence begets bias

- Low powered studies systematically report larger effects. Why?
  - How else will they be statistical significant? And, they tend to be reported as statistically sig.
  - Among the 31 areas of research that do not contain a single adequately powered estimate, 44% report statistically significant results.
- The Paldam Principle: Divide all reported estimates by 2—in the *AER* or anywhere.

Let’s ‘WAAP’ this Bias in it tracks

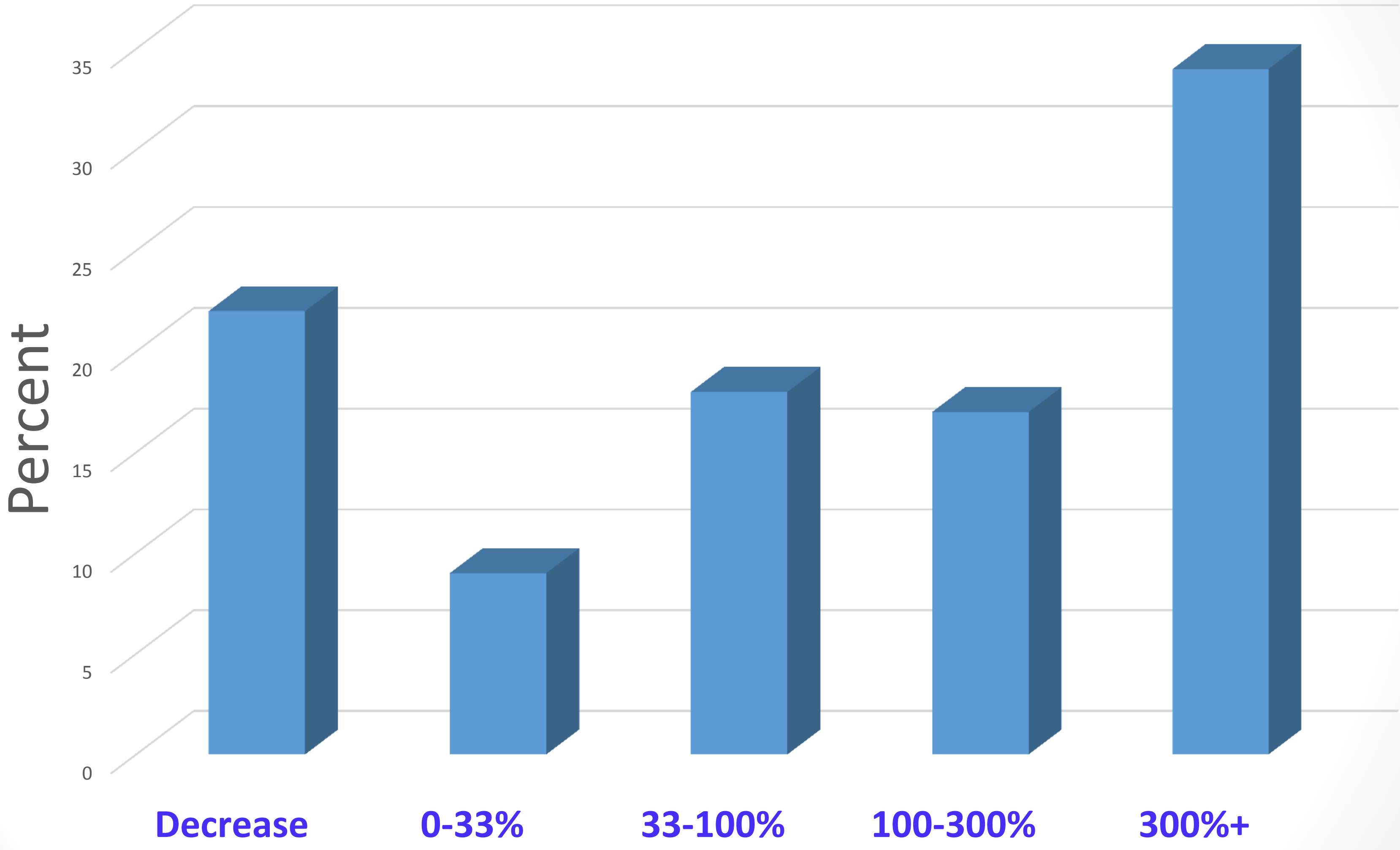
WAAP:

- is the weighted average of only the adequately powered estimates.
- uses our unrestricted WLS weighted average—the exact same point estimate as fixed-effect.
- is onomatopoeia.
- offers a conservative assessment of bias; it offers an empirical lower bound for bias in economics.

Assessing Research Inflation

- **Research Inflation** is the difference between the average reported effect and some proxy for 'true' effect. It calculates bias empirically.
- We use **WAAP** and **PET-PEESE** as the proxies for the 'true' empirical effect.
- **WAAP** cannot further bias the reported estimates. It uses only the objectively best (most powerful) estimates.
- Those researchers who are skeptical of **PET-PEESE** cannot object to **WAAP**.

Figure 3: Research Inflation, *WAAP*



Deflation | Percent Increase

How Biased is Economics?

- **The Paldam Principle is confirmed**!
- The median research inflation is just over 100%—that is, typically economics is exaggerated by a factor of 2 or more.
- Stanley Corollary: OK, 2 is about right, but many areas are even more biased.
- At least 1/3 of economics is exaggerated by a factor of 4 or more.
- The 1/3rd solution. . . . ?

In sum, this meta-meta-analysis has the weight of 159 Paldams!

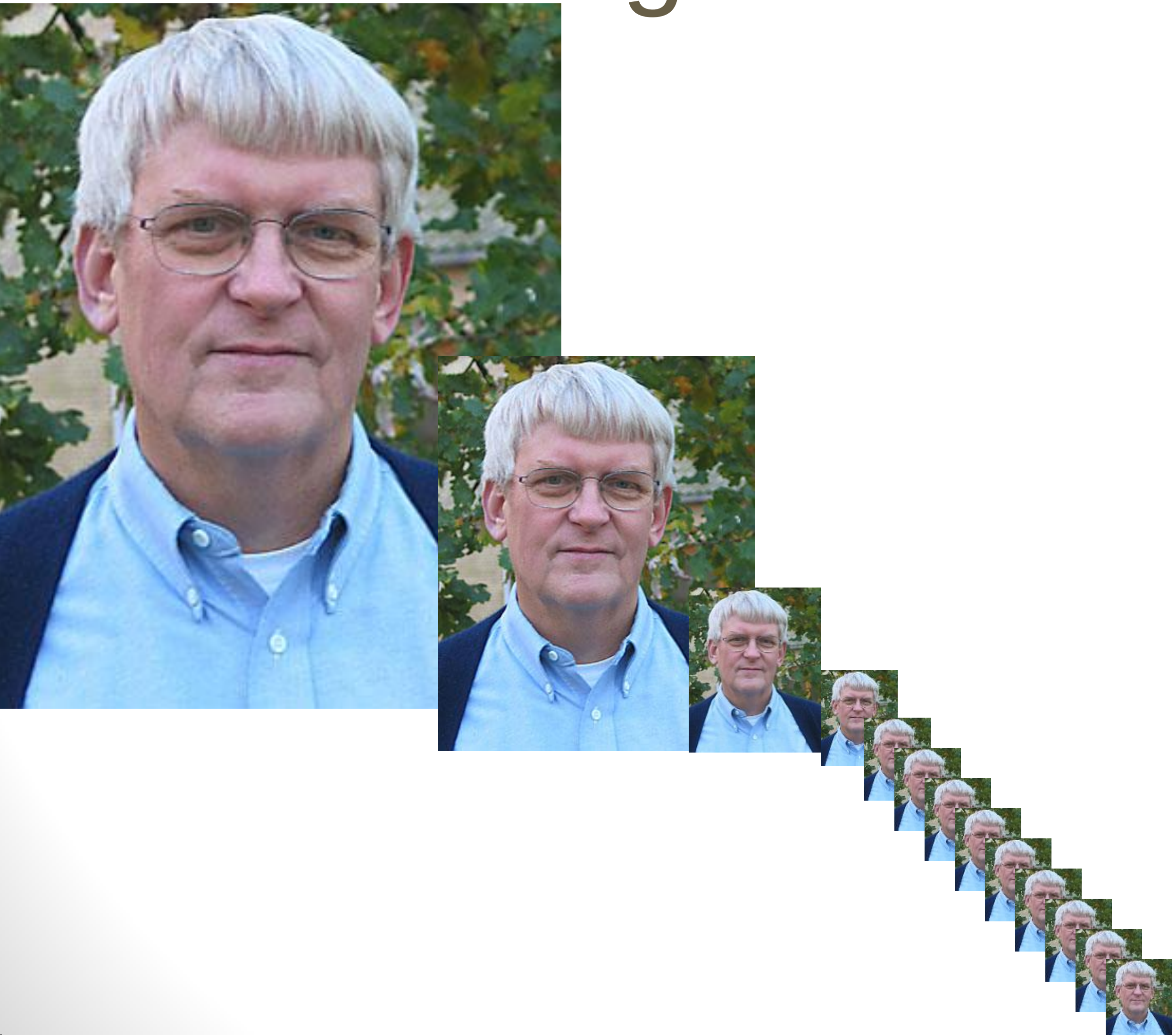
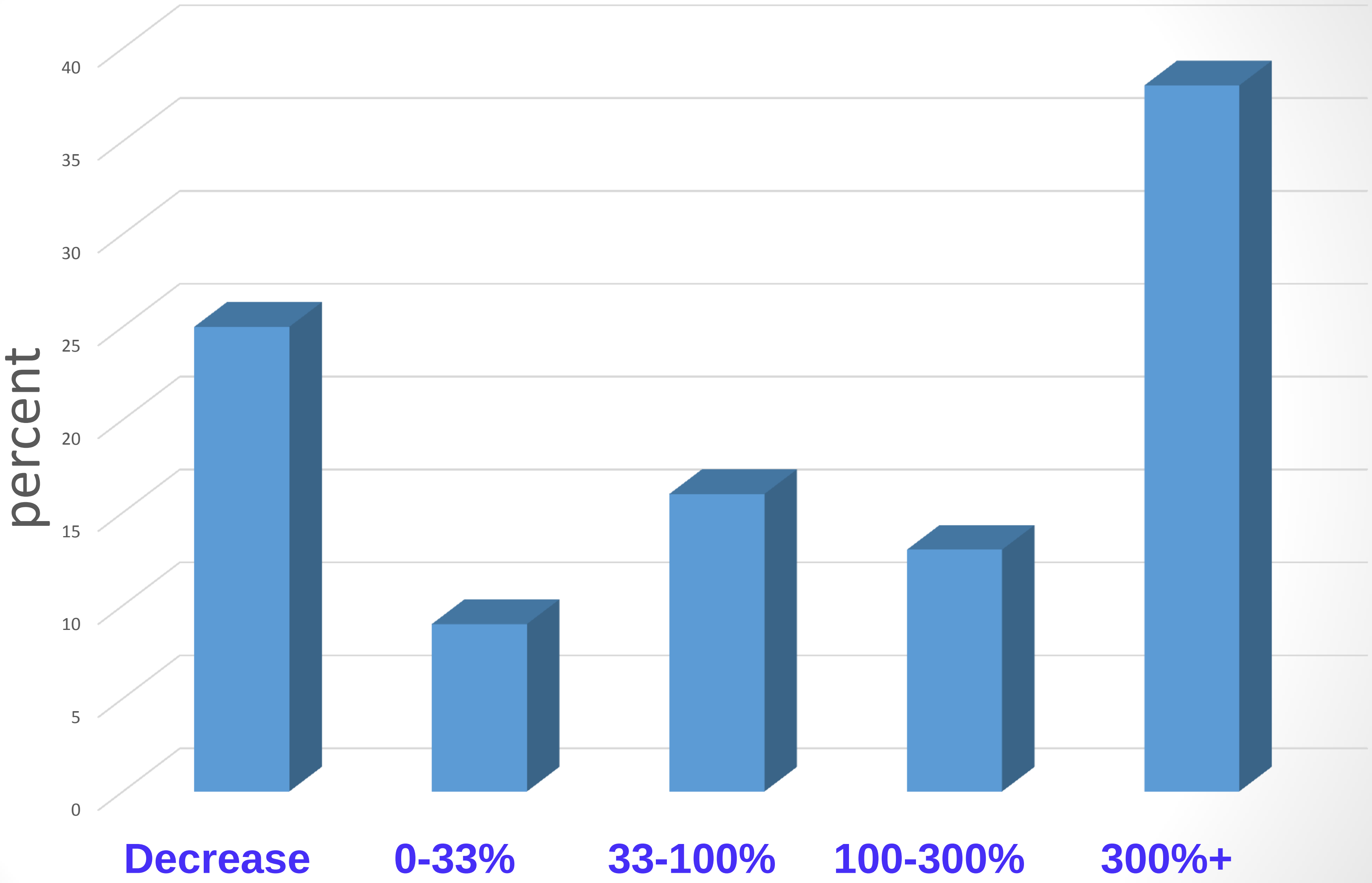
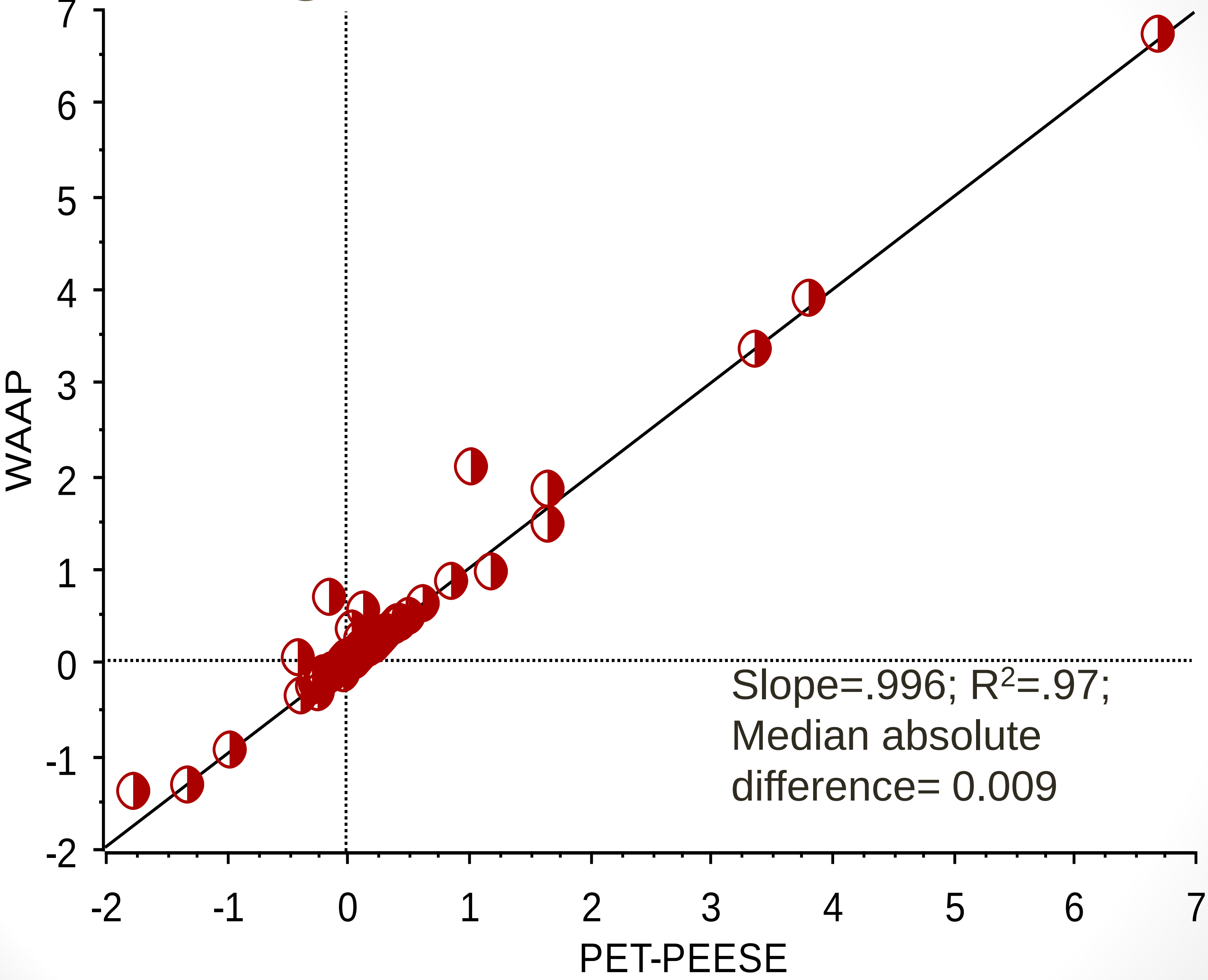


Figure 4: Research Inflation, *PET-PEESE*



Deflation | Percent Increase

Predicting WAAP from PET-PEESE



Slope=.996; R^2=.97;
Median absolute
difference= 0.009

Implications

- It’s **bad**!
- It’s REALLY **bad**.
- Good news: Other disciplines are also **bad**!
  - Psychology routinely has low power.
    - Nosek, B. A. et al. ***Science*** Aug. 2015 found that effect size shrunk in half when 100 psych experiments were replicated. **+ one Pladam**
  - Among 14,886 meta-analyses in the Cochrane Database of Systematic Reviews, the median power to detect a medium-size effect is 13%--(Turner et al., 2013).

What to do about low power and large biases in Economics?

- Don’t worry; be happy.
- Heck, medical research is nearly as bad.
- Change the research incentive structure:
  - Little value for publication in high impact journals
  - Much more for a study that has been replicated
  - Credit for sharing data and codes.
- Require all empirical studies to report power from:
  - a previous meta-analysis or
  - a systematic and realistic assessment of what past studies have found.