The need for evidence-based decision making and science reform

heil og sæl

takk for at du spurte meg

@metanutter

Gavin.Stewart@Newcastle.ac.uk

Why do we need evidence?

- The challenge of feeding nine billion people
  - No more land, climate change, increasing variability

Science 327, 812 (2010)

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

But lots of “evidence” is wrong

What is evidence….Is expert judgement evidence?

How often do experts make the right predictions?

All evidence needs value judgements to assess its strength.

Ioannidis JPA (2005)
Why Most Published Research Findings Are **False**. *PLOS Medicine*

What does our evidence look like?

- The **replication crisis**

Schooler, J. W. (2014). "Metascience could rescue the 'replication crisis'". *Nature*. 515 (7525): 9.

Empirical evidence

Domain	Findings	Sources
Medicine	Out of 49 highly cited papers, 45 claimed that studied therapy was effective. Of these studies, 16% were contradicted by subsequent studies, 16% had found stronger effects than did subsequent studies, 44% were replicated, and 24% remained largely unchallenged. 11% of pre-clinical cancer studies were replicable	Ioannidis JA (13 July 2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. **294** (2): 218–228. Begley, CG., and Lee ME., (2012) Drug Development: Raise Standards for Preclinical Cancer Research, *Nature*. **483**, 531–533.
Psychology	Out of 100 studies from high-ranking journals only 36% had significant findings (*p* value below .05) compared to 97% of the original studies. The mean effect size in the replications was approximately half the magnitude of the effects reported in the original studies. Questionable research practices (QRPs) have been identified as common in the field (majority of 2000 scientists confess to at least one of: e.g. selective reporting, p-hacking, nonpublication of data, post-hoc storytelling (framing exploratory analyses as confirmatory analyses), manipulation of outliers.	Collaboration, Open Science (2015). "Estimating the reproducibility of psychological science". *Science*. **349** (6251): aac4716. Leslie JK.; Loewenstein, GP, Drazen (2012). "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling". *Psychological Science*. **23** (5): 524–532

The dance of the P values

Strength of evidence
P<0.001
P<0.01
P<0.05
P 0.05 to ?
P>0.1

The classical P value: The probability of observing data at least as extreme as the actual data given infinite observations.... assuming the null hypothesis to be true

https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

The dance of the P values

Strength of evidence	Significance language
P<0.001	Very highly Significant
P<0.01	Highly significant
P<0.05	Significant
P 0.05 to ?	Approaching Significant
P>0.1	Non-significant

https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

The dance of the P values

Strength of evidence	Significance language	Suggests Truth
P<0.001	Very highly Significant	There is definitely an effect
P<0.01	Highly significant	There is an effect
P<0.05	Significant	Most likely there is an effect
P 0.05 to ?	Approaching Significant	Almost? Probably? (but low power)
P>0.1	Non-significant	No effect?

https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

The dance of the P values

Strength of evidence	Significance language	Suggests Truth	Evokes emotion
P<0.001	Very highly Significant	There is definitely an effect	Elation Exuberance Smugness?
P<0.01	Highly significant	There is an effect	Dancing, Drinking
P<0.05	Significant	Most likely there is an effect	Relief Cheerfulness
P 0.05 to ?	Approaching Significant	Almost? Probably? (but low power)	Frustration (if only)
P>0.1	Non-significant	No effect?	Despair, depression

https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

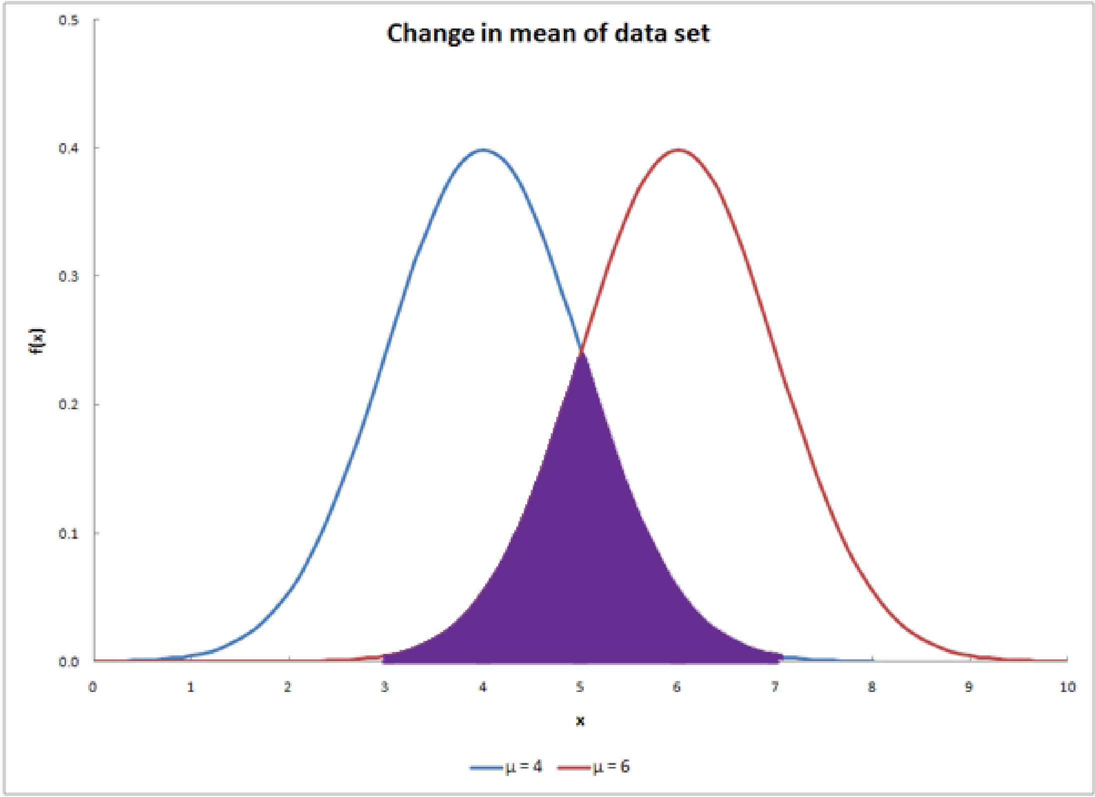
The dance of the P values

Strength of evidence	Significance language	Suggests Truth	Evokes emotion	Implications
P<0.001	Very highly Significant	There is definitely an effect	Elation Exuberance Smugness?	Nobel Prize Tenure Research Grant
P<0.01	Highly significant	There is an effect	Dancing, Drinking	**** publication PhD
P<0.05	Significant	Most likely there is an effect	Relief Cheerfulness	*** publication
P 0.05 to ?	Approaching Significant	Almost? Probably? (but low power)	Frustration (if only)	Stress leave counselling
P>0.1	Non-significant	No effect?	Despair, depression	Reconsider life goals

https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

The Dance of the P values

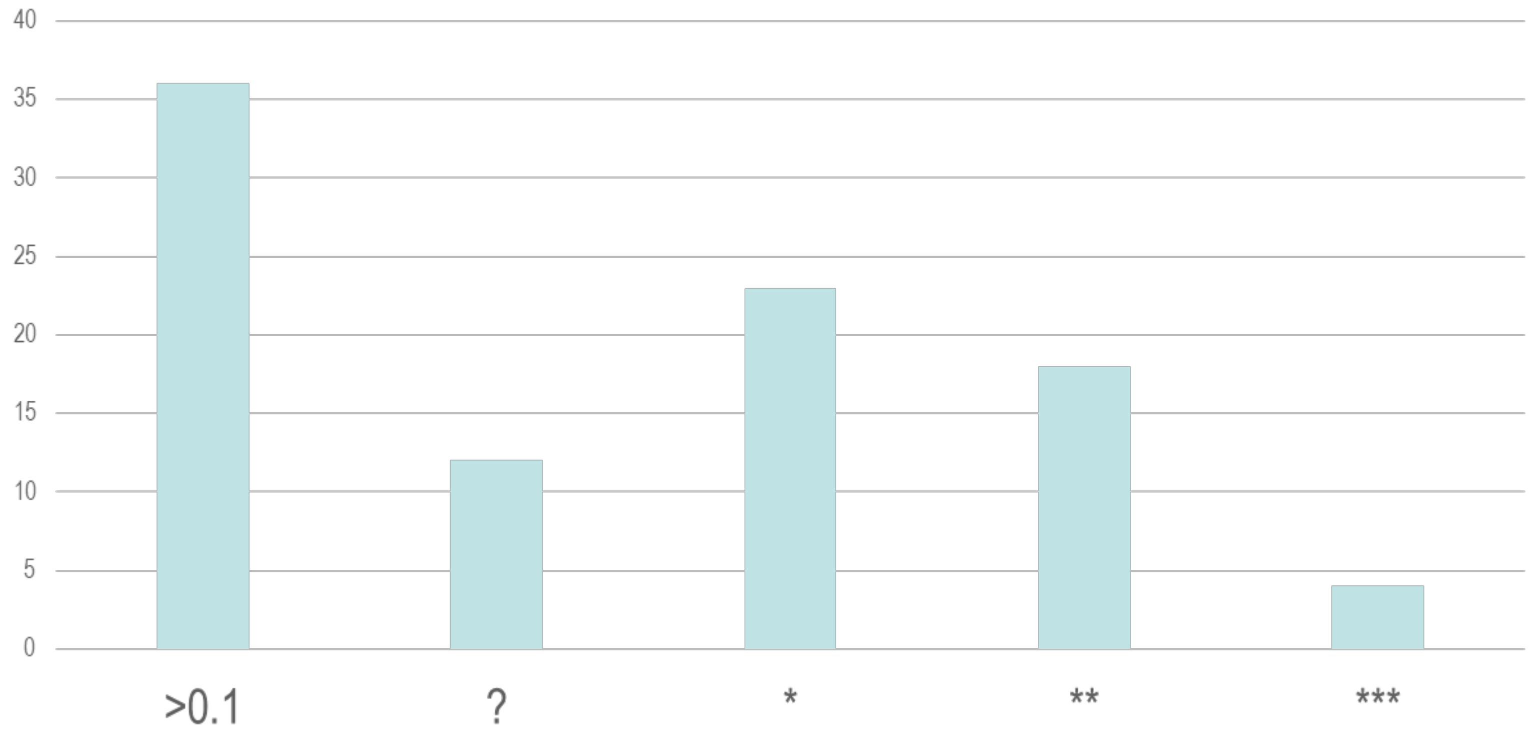
- If P values are meaningful and represent the truth they should replicate...
- Let’s run a simulation to see if they do…



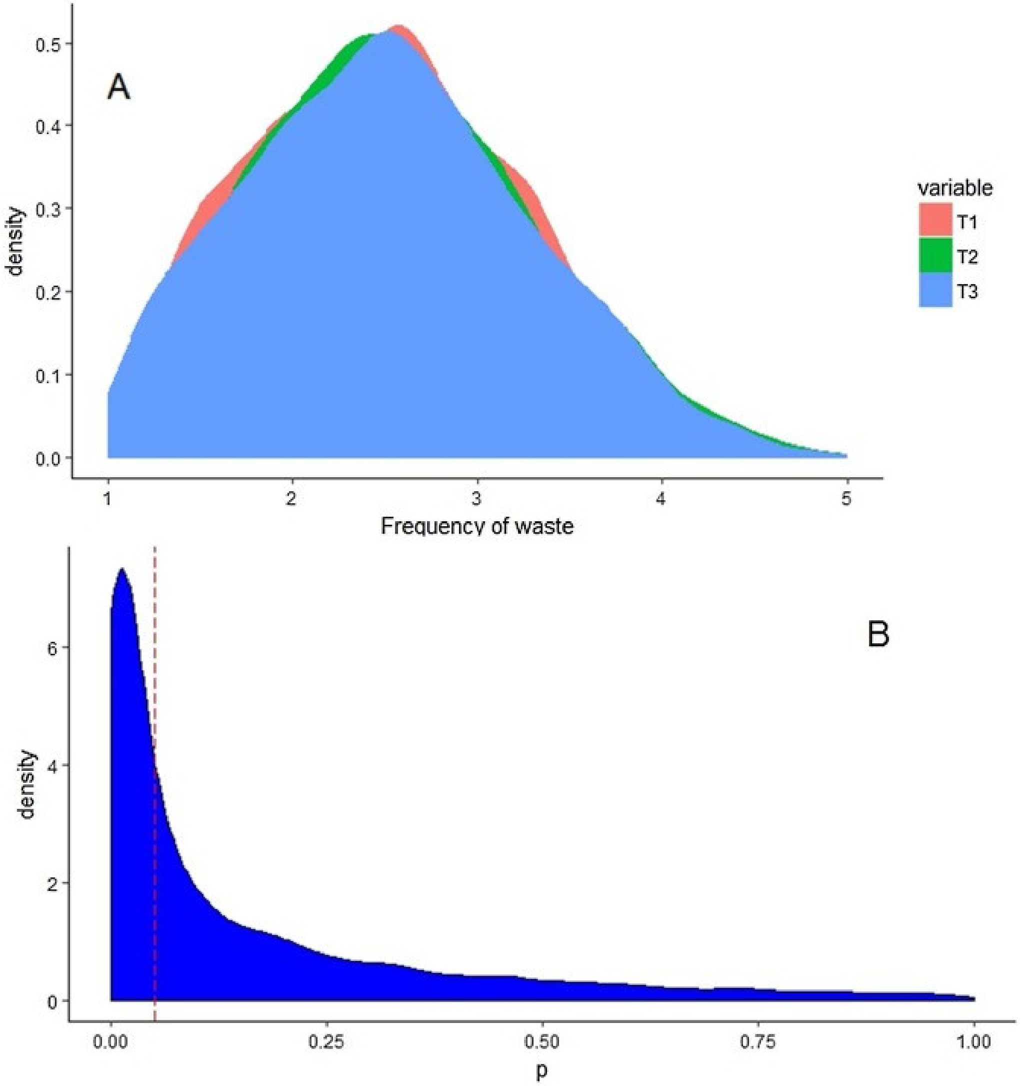
https://www.youtube.com/watch?v=5OL1RqHrZQ8

Dance of the P values

- P values do not replicate
- (Over)reliance on P values has serious consequences for the rigour of our science…



A real example where p values mislead..



Grainger MJ, Stewart GB. **The jury is still out on social media as a tool for reducing food waste a response to Young et al. (2017)**. *Resources, Conservation and Recycling* 2017, **122**, 407-410.

Publication bias

- Publication bias refers to bias that occurs when research found in the published literature is **systematically unrepresentative** of the population of studies (Rothstein et al., 2005)

- On average published studies have a larger mean effect size than unpublished studies, providing evidence for a publication bias (Lipsey and Wilson 1993)

- Also referred to as the **'file drawer' problem**:

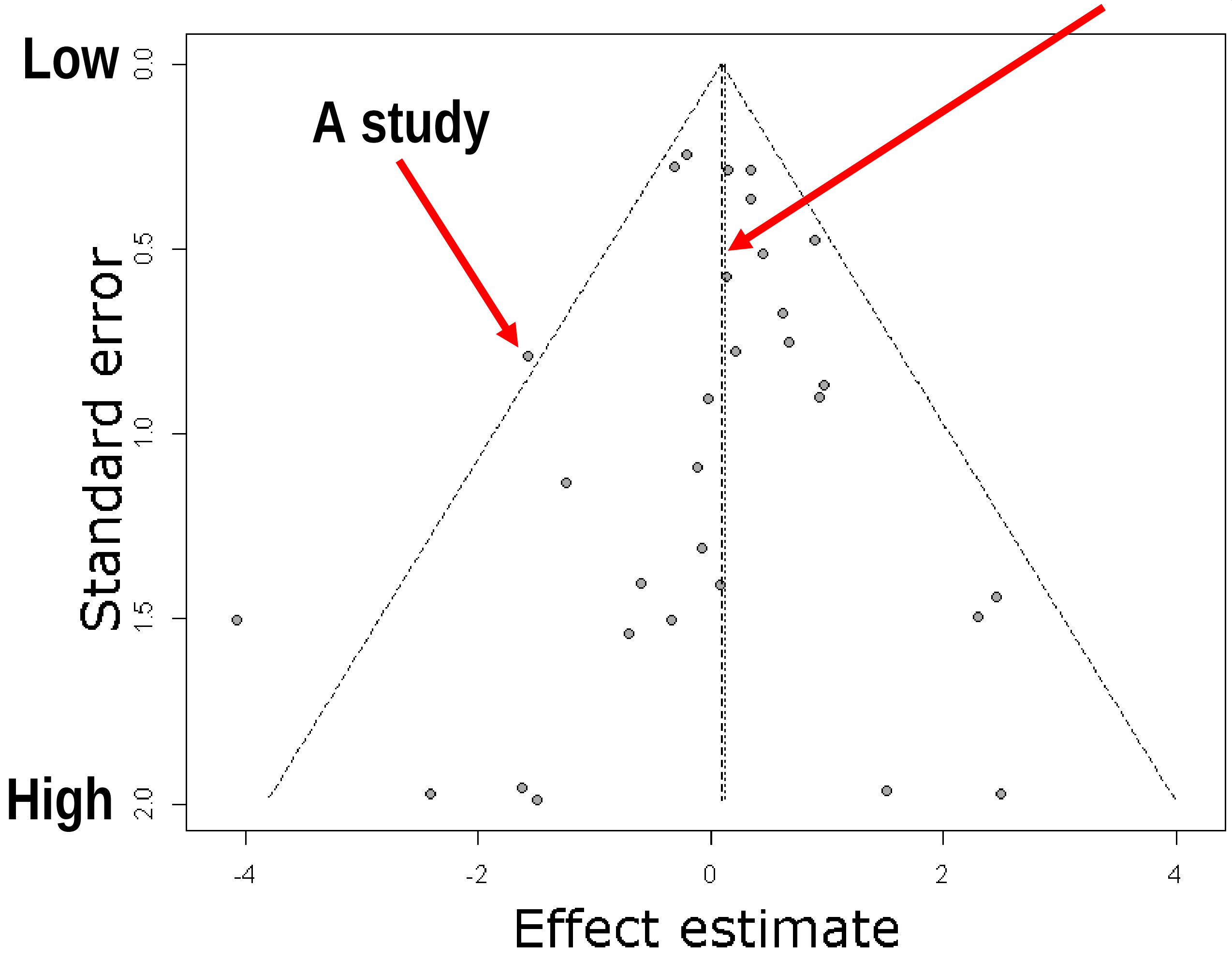
"…journals are filled with the 5% of studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g. p < 0.05) results" (Rosenthal, 1979)

- Well-documented in different fields of research (biomedicine, public health, education, crime & justice, social welfare, ecology & evolution).

Rothstein, H. R., Sutton, A. J., & Borenstein, M. L. (Eds). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

The funnel plot

True effect
from meta-analysis



A study

Large studies
close to true effect

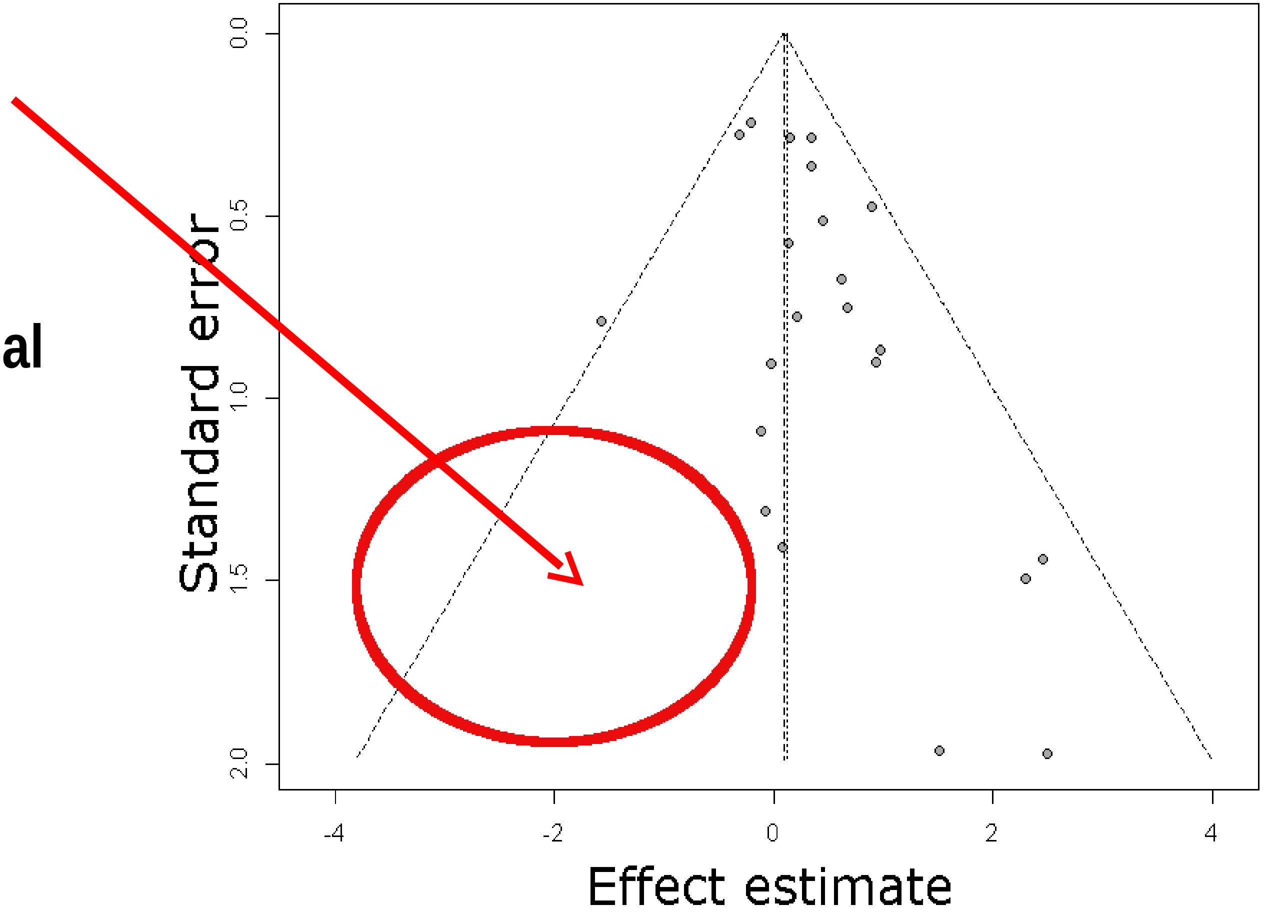
95% of studies should
be in the "funnel"

Small studies
more variable

Now with added publication bias

Studies missing from lower corner of funnel

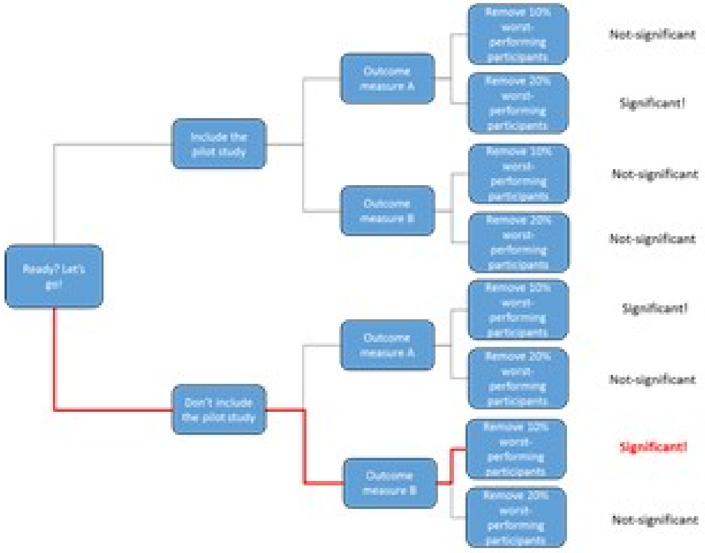
Funnel is not symmetrical



Sterne J *et al.* (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 343, d4002.

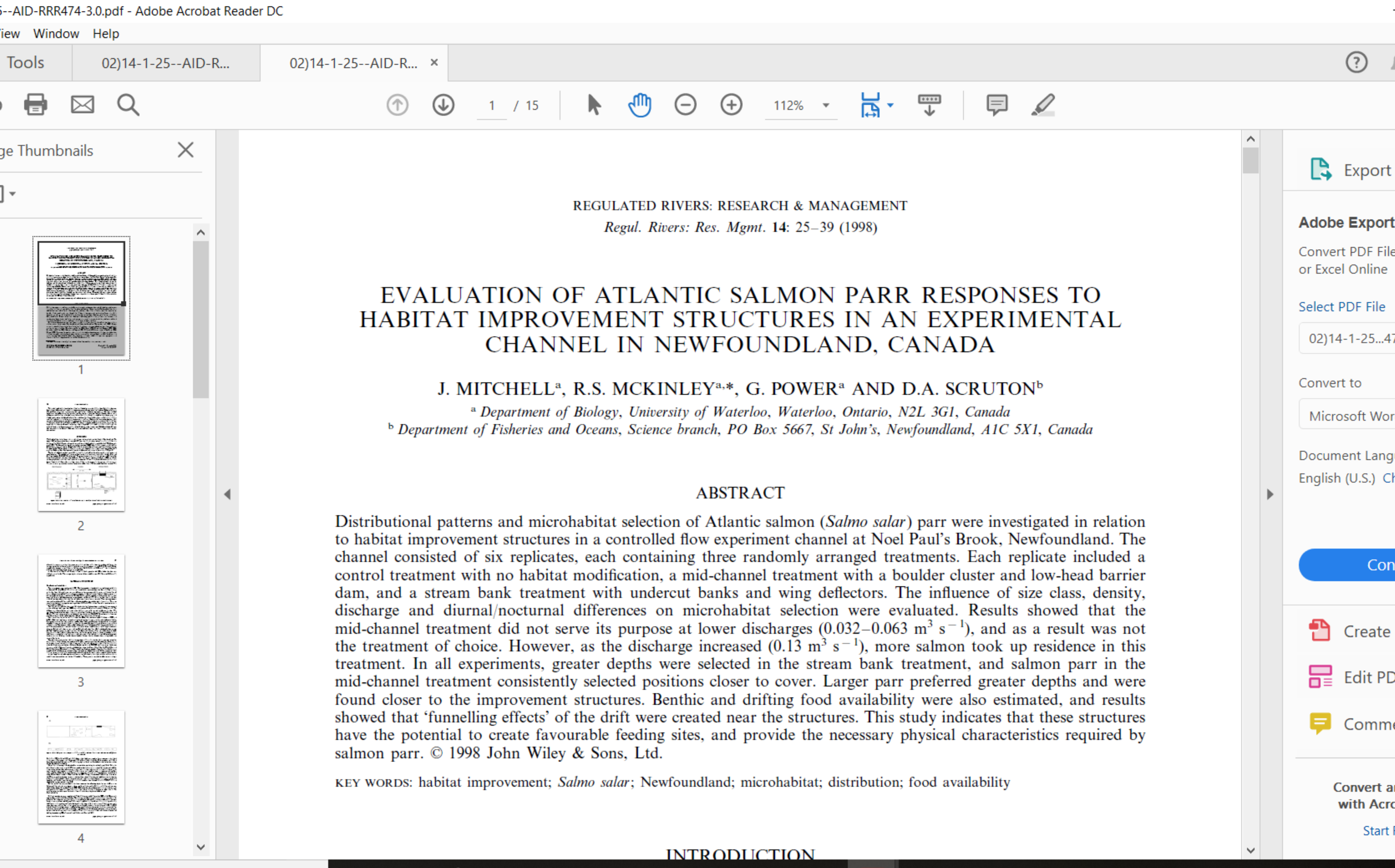
Reporting and researcher degrees of freedom

- Do lots of things in different ways…and consciously or unconsciously introduce bias with selective reporting
- Develop an SEM with two different structures, split the data into male and female, analyse complete cases and imputed data…report only selected results (and worse selected methods)
- And just bad reporting of important information


Ready? Let's go!
Include the pilot study
Don't include the pilot study
Outcome measure A
Outcome measure B
Outcome measure A
Outcome measure B
Remove 10% worst-performing participants
Not-significant
Remove 20% worst-performing participants
Significant!
Remove 10% worst-performing participants
Not-significant
Remove 20% worst-performing participants
Not-significant
Remove 10% worst-performing participants
Significant!
Remove 20% worst-performing participants
Not-significant
Remove 10% worst-performing participants
Significant!
Remove 20% worst-performing participants
Not-significant

“Yes, Significant!”

A real example of researcher degrees of freedom



REGULATED RIVERS: RESEARCH & MANAGEMENT
Regul. Rivers: Res. Mgmt. **14**: 25–39 (1998)

EVALUATION OF ATLANTIC SALMON PARR RESPONSES TO HABITAT IMPROVEMENT STRUCTURES IN AN EXPERIMENTAL CHANNEL IN NEWFOUNDLAND, CANADA

J. MITCHELL[a], R.S. MCKINLEY[a,*], G. POWER[a] AND D.A. SCRUTON[b]
[a] *Department of Biology, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*
[b] *Department of Fisheries and Oceans, Science branch, PO Box 5667, St John's, Newfoundland, A1C 5X1, Canada*

ABSTRACT

Distributional patterns and microhabitat selection of Atlantic salmon (*Salmo salar*) parr were investigated in relation to habitat improvement structures in a controlled flow experiment channel at Noel Paul's Brook, Newfoundland. The channel consisted of six replicates, each containing three randomly arranged treatments. Each replicate included a control treatment with no habitat modification, a mid-channel treatment with a boulder cluster and low-head barrier dam, and a stream bank treatment with undercut banks and wing deflectors. The influence of size class, density, discharge and diurnal/nocturnal differences on microhabitat selection were evaluated. Results showed that the mid-channel treatment did not serve its purpose at lower discharges (0.032–0.063 $m^3 s^{-1}$), and as a result was not the treatment of choice. However, as the discharge increased (0.13 $m^3 s^{-1}$), more salmon took up residence in this treatment. In all experiments, greater depths were selected in the stream bank treatment, and salmon parr in the mid-channel treatment consistently selected positions closer to cover. Larger parr preferred greater depths and were found closer to the improvement structures. Benthic and drifting food availability were also estimated, and results showed that 'funnelling effects' of the drift were created near the structures. This study indicates that these structures have the potential to create favourable feeding sites, and provide the necessary physical characteristics required by salmon parr. © 1998 John Wiley & Sons, Ltd.

KEY WORDS: habitat improvement; *Salmo salar*; Newfoundland; microhabitat; distribution; food availability

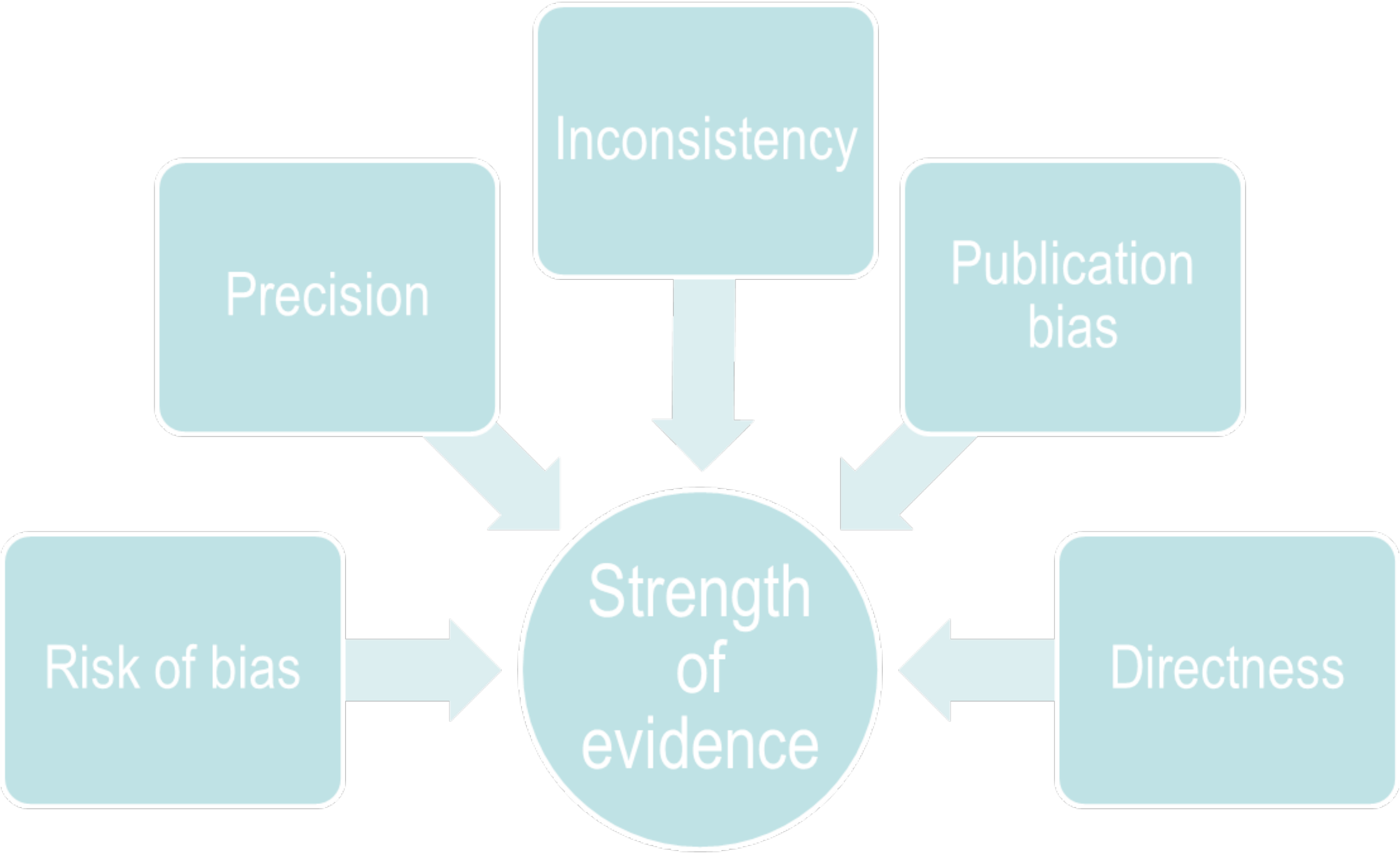
INTRODUCTION

Novelty and theory

- Good research must be novel with sound theoretical underpinnings?

- Or is causation more important?

Good research updates our belief about evidence



Stewart G, Higgins J, Schunneman H, Meader N. (2015) The use of Bayesian Networks to assess the quality of evidence from research synthesis. *PLoS ONE* 10(4)

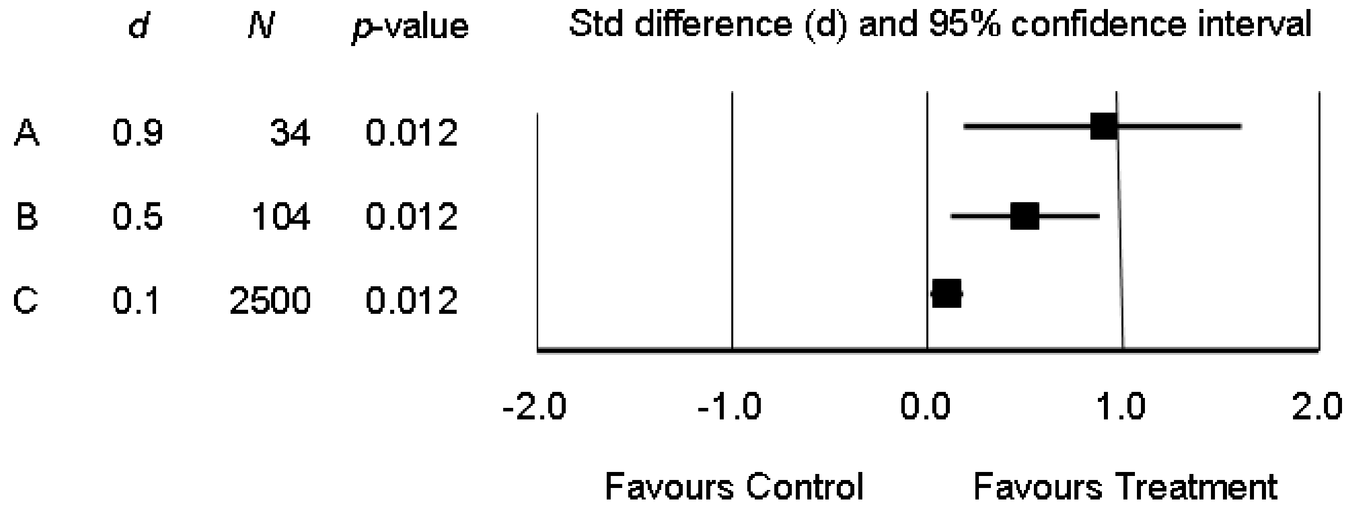
Summary to date

- We’re **BAD**

- Over(reliance on p values)
- Publication bias
- Selective reporting and story telling
- Inappropriate emphasis on novelty with failure to standardise measurements
- Fail to consider cumulative evidence appropriately
- Poor reporting *

Solutions 1: P values

- Report and interpret effect sizes and confidence intervals (they convey much more information than p values)



	d	*N*	*p*-value
A	0.9	34	0.012
B	0.5	104	0.012
C	0.1	2500	0.012

- Establish univ https://www.equator-network.org

- Some advocacy for banning p values altogether

Nuzzo R (2014) Nature 506:150-152

Solution 2: Publication Bias

- Pre-registration
- TOP guidelines
  - Pre-registered
  - Open Data
  - Open Methods

CENTER FOR OPEN SCIENCE

Solution 3: selective and poor reporting

- See previous:
  - Less reliance on p values
  - Adherence to reporting guidelines
  - Pre-registration, open data, open methods

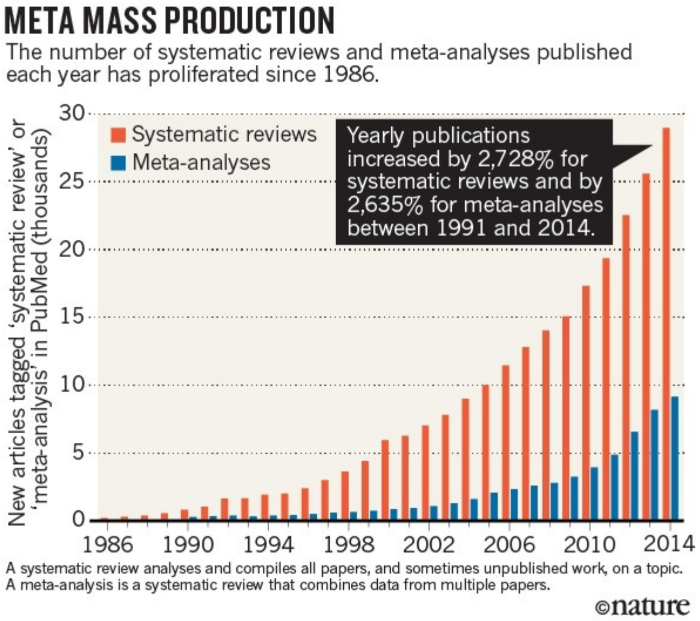
What is your n?

Solution 4: considering the cumulative evidence

- More high quality evidence synthesis
  - Inform policy without the hype
  - Exposure to deficiencies in current evidence

- Strength of evidence rather than novelty

- Systems approach to funding
  - Informed by ES and informing ES
  - Common outcomes rather than novelty



META MASS PRODUCTION

The number of systematic reviews and meta-analyses published each year has proliferated since 1986.

Yearly publications increased by 2,728% for systematic reviews and by 2,635% for meta-analyses between 1991 and 2014.

A systematic review analyses and compiles all papers, and sometimes unpublished work, on a topic. A meta-analysis is a systematic review that combines data from multiple papers.

©nature

New articles tagged 'systematic review' or 'meta-analysis' in PubMed (thousands)
30
25
20
15
10
5
0
🟥 Systematic reviews
🟦 Meta-analyses
1986
1990
1994
1998
2002
2006
2010
2014

The Milbank Quarterly, Vol. 94, No. 3, 2016 (pp. 485-514)

Solution 5: more meta-science

- What is a large effect in discipline X
- How large is the effect in the first study compared to the largest study in area Y
- How many studies are wrong because of hacking or harking?

Acknowledgements

SRSM

Society for Research
Synthesis Methodology

CampbellCollaboration

Cochrane