"'Repligate': reproducibility in statistical studies. What does it mean and in what sense does it matter?"

Stephen Senn

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Acknowledgements

Thanks to Rogier Kievit and Richard Morey for the invitation and to APS for support

This work is partly supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDEAL"

Outline

- The crisis of replication?
- A brief history of P-values
- What are we looking for in replication?
- Empirical evidence
- Conclusions

The Crisis of Replication

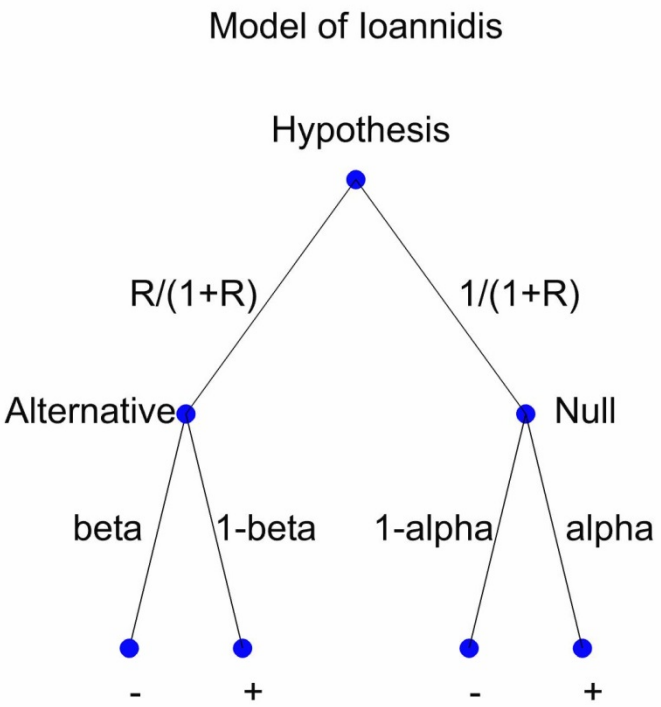
In countless tweets….The “replication police” were described as “shameless little bullies,” “self-righteous, self-appointed sheriffs” engaged in a process “clearly not designed to find truth,” “second stringers” who were incapable of making novel contributions of their own to the literature, and—most succinctly—“assholes.”

Why Psychologists’ Food Fight Matters
“Important findings” haven’t been replicated, and science may have to change its ways.

By Michelle N. Meyer and Christopher Chabris , *Science*

Ioannidis (2005)

- Claimed that most published research findings are wrong
  - By finding he means a 'positive' result
- 2764 citations by 18 May 2015 according to Google Scholar



Colquhoun's Criticisms

One must admit, however reluctantly, that despite the huge contributions that Ronald Fisher made to statistics, there is an element of truth in the conclusion of a perspicacious journalist:

> The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug. Robert Matthews [21] *Sunday Telegraph*, 13 September 1998.

"If you want to avoid making a fool of yourself very often, do not regard anything greater than p <0.001 as a demonstration that you have discovered something. Or, slightly less stringently, use a three-sigma rule."

Royal Society Open Science 2014

A Common Story

- Scientists were treading the path of Bayesian reason
- Along came RA Fisher and persuaded them into a path of P-value madness
- This is responsible for a lot of unrepeatable nonsense
- We need to return them to the path of Bayesian virtue
- In fact the history is not like this and understanding this is a key to understanding the problem

From the table the probability is ·9985 or the odds are about 666 to 1 that 2 is the better soporific.

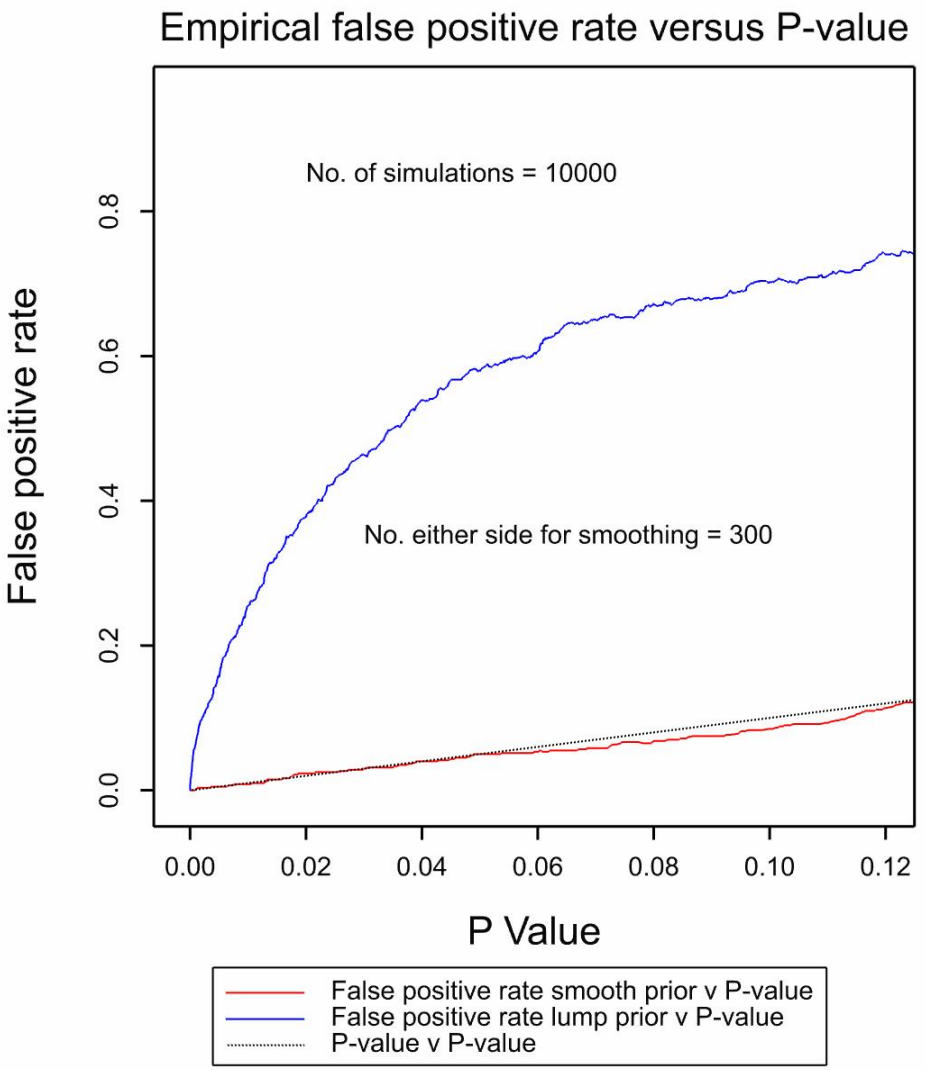
Student, The Probable Error of a Mean, *Biometrika*, 1908, P21

For $n = 9$, only one value in a hundred will exceed 3·250 by chance, so that the difference between the results is clearly significant.

Fisher, *Statistical Methods for Research Workers*, 1925

The real history

- Scientists before Fisher were using tail area probabilities to calculate posterior probabilities
- Fisher pointed out that this interpretation was unsafe and offered a *more conservative* one
- Jeffreys, influenced by CD Broad's criticism, was unsatisfied with the Laplacian framework and used a lump prior probability on a point hypothesis being true
- It is Bayesian Jeffreys versus Bayesian Laplace that makes the dramatic difference, not frequentist Fisher versus Bayesian Laplace


Empirical false positive rate versus P-value
No. of simulations = 10000
No. either side for smoothing = 300
False positive rate
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.06
0.08
0.10
0.12
P Value
False positive rate smooth prior v P-value
False positive rate lump prior v P-value
P-value v P-value

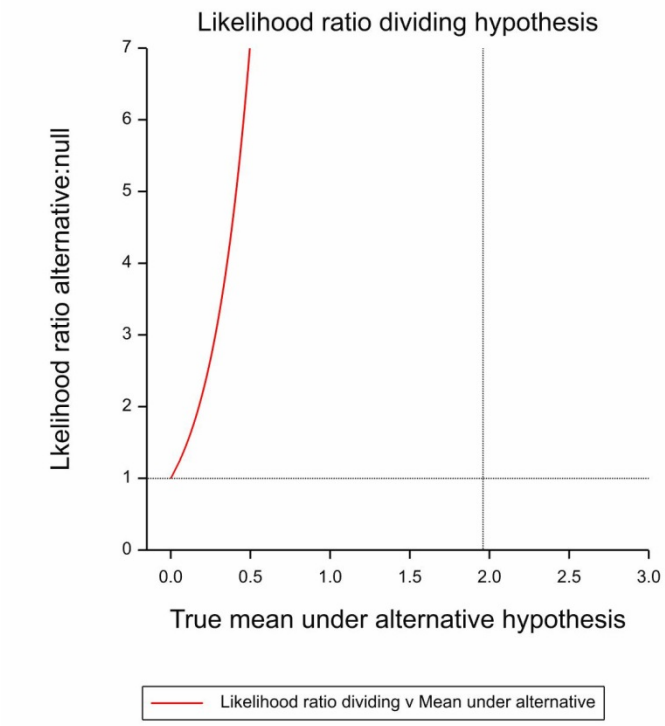
Why the difference?

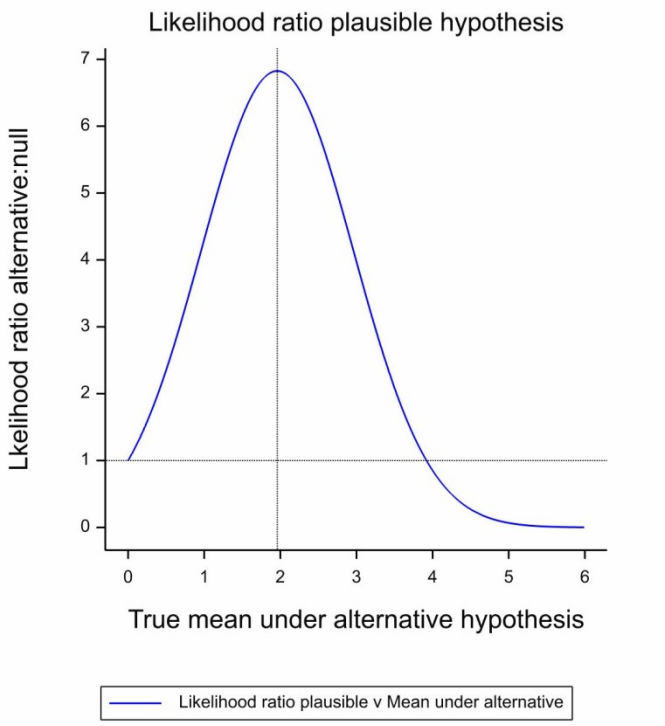
- Imagine a point estimate of two standard errors
- Now consider the likelihood ratio for a given value of the parameter, δ under the alternative to one under the null
  - *Dividing hypothesis (smooth prior)* for any given value $\delta = \delta'$ compare to $\delta = -\delta'$
  - Plausible hypothesis (lump prior) for any given value $\delta = \delta'$ compare to $\delta = 0$

A speculation of mine

- Scientists had noticed that for dividing hypotheses they could get ‘significance’ rather easily
  - The result is the 1/20 rule
- However when deciding to choose a new parameter or not in terms of probability it is 50% not 5% that is relevant
- This explains the baffling finding that significance test are actually *more* conservative than AIC (and sometimes than BIC)

The situations compared





Goodman's Criticism

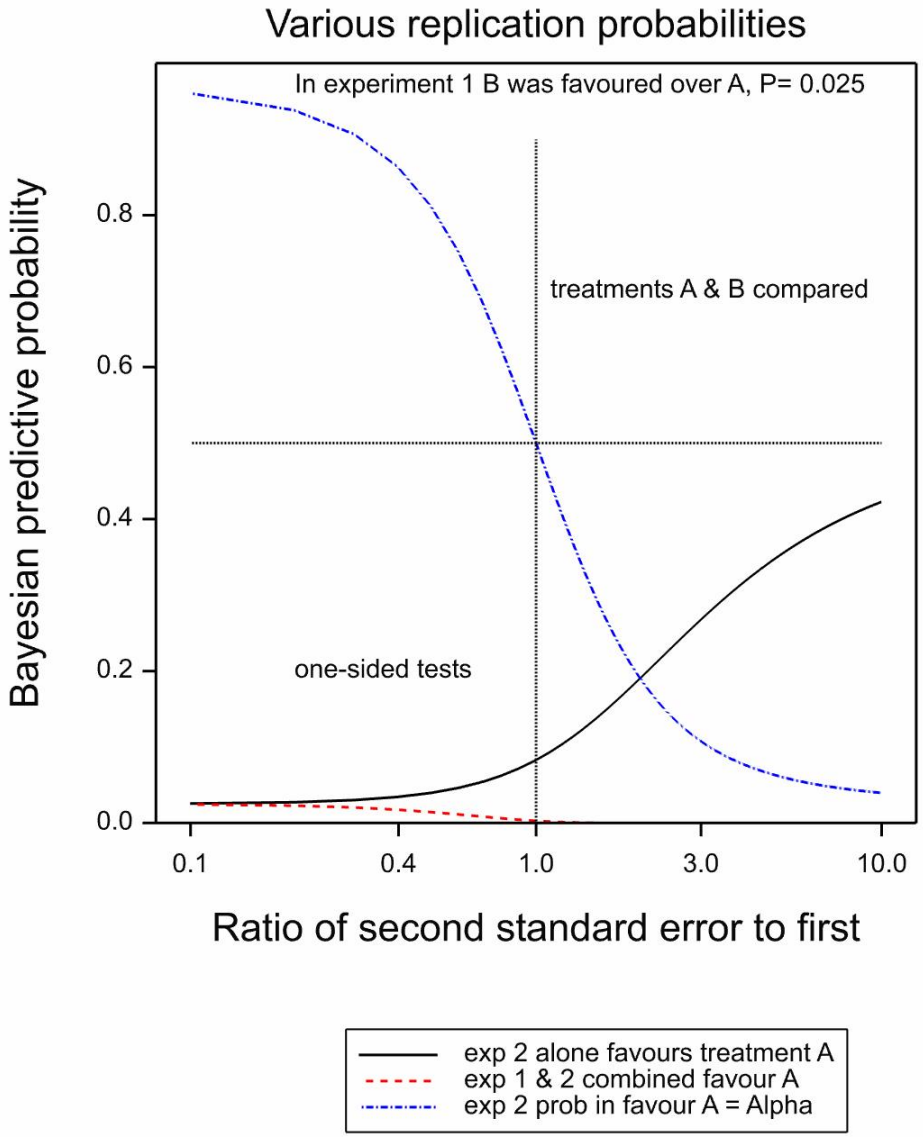
- What is the probability of repeating a result that is just significant at the 5% level (p=0.05)?
- Answer 50%
  - If true difference is observed difference
  - If uninformative prior for true treatment effect
- Therefore P-values are unreliable as inferential aids

Sauce for the Goose and Sauce for the Gander

- This property is shared by Bayesian statements
  - It follows from the Martingale property of Bayesian forecasts
- Hence, either
  - The property is undesirable and hence is a criticism of Bayesian methods also
  - Or it is desirable and is a point in favour of frequentist methods

Three Possible Questions

- Q1 What is the probability that in a future experiment, taking that experiment's results *alone*, the *estimate* for B would after all be worse than that for A?
- Q2 What is the probability, having conducted this experiment, and *pooled* its results with the current one, we would show that the *estimate* for B was, after all, worse than that for A?
- Q3 What is the probability that having conducted a future experiment and then calculated a Bayesian posterior using a uniform prior and the results of this second experiment *alone*, the *probability* that B would be worse than A would be less than or equal to 0.05?


Various replication probabilities
In experiment 1 B was favoured over A, P= 0.025
treatments A & B compared
one-sided tests
Bayesian predictive probability
0.0
0.2
0.4
0.6
0.8
0.1
0.4
1.0
3.0
10.0
Ratio of second standard error to first
exp 2 alone favours treatment A
exp 1 & 2 combined favour A
exp 2 prob in favour A = Alpha

Why Goodman's Criticism is Irrelevant

"It would be absurd if our inferences about the world, having just completed a clinical trial, were *necessarily* dependent on assuming the following. 1. We are now going to repeat this experiment. 2. We are going to repeat it only once. 3. It must be exactly the same size as the experiment we have just run. 4. The inferential meaning of the experiment we have just run is the extent to which it predicts this second experiment."

Senn, 2002

A Paradox of Bayesian Significance Tests

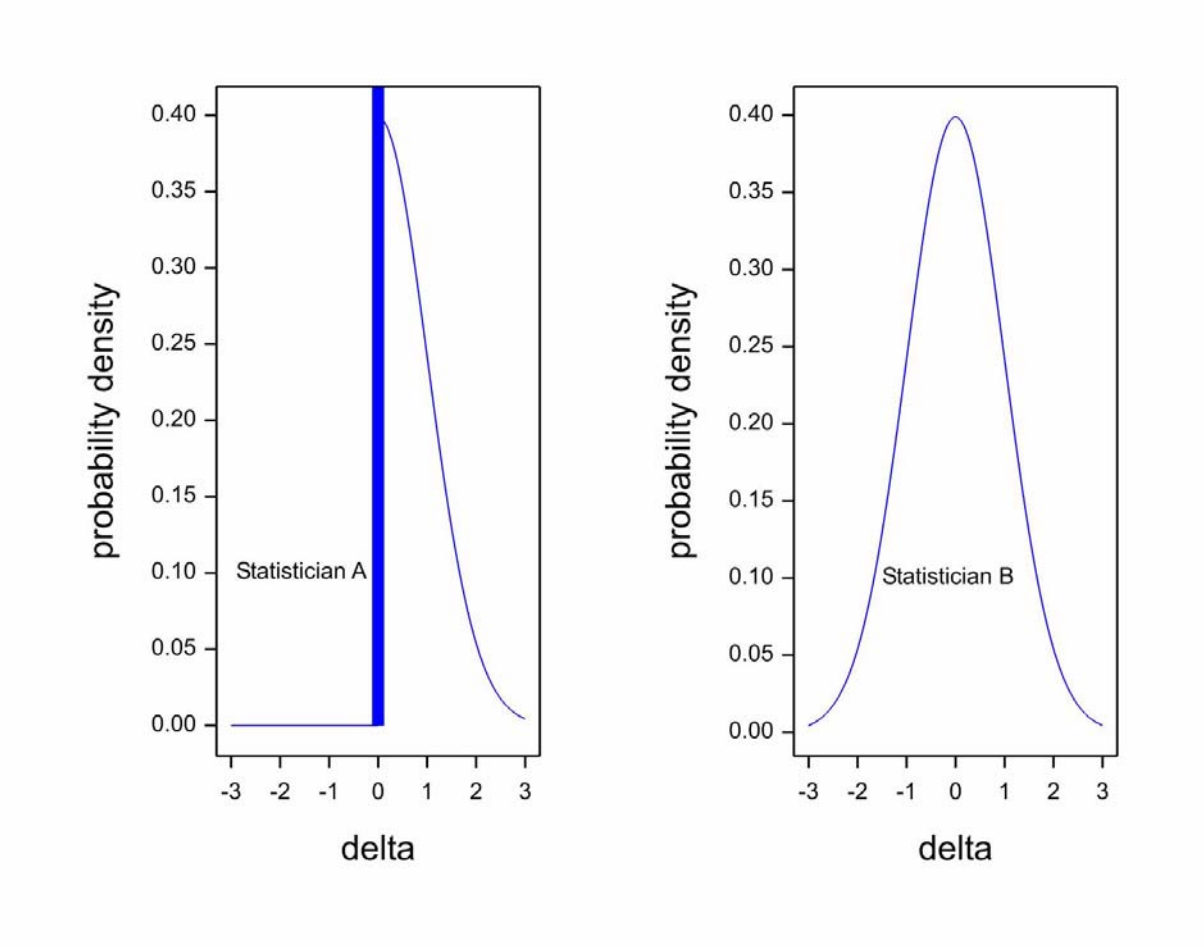
Two scientists start with the same probability 0.5 that a drug is effective.

Given that it is effective they have the same prior for how effective it is.

If it is not effective A believes that it will be useless but B believes that it may be harmful.

Having seem the same data B now believes that it is useful with probability 0.95 and A believes that it is useless with probability 0.95.

A Tale of Two priors



In Other Words

The probability is 0.95

And the probability is also 0.05

Both of these probabilities can be simultaneously true.

NB This is *not* illogical but it is illogical to regard this sort of thing as proving that p-values are illogical

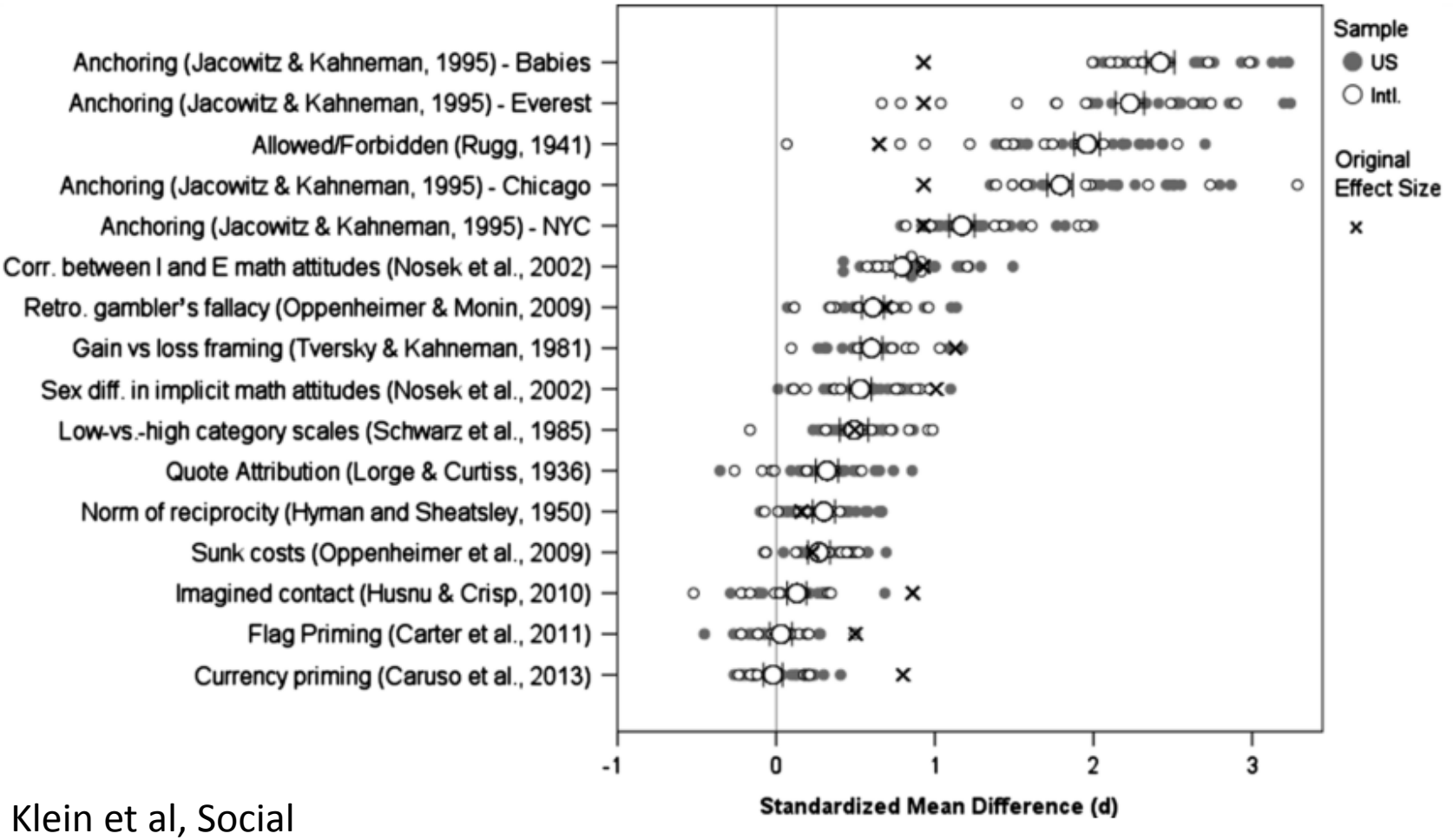
'…would require that a procedure is dismissed because, when combined with information which it doesn't require and which may not exist, it disagrees with a procedure that disagrees with itself'

Senn, 2000

Are most research findings false?

- A dram of data is worth a pint of pontification
- Two interesting studies recently
  - The many labs replications project
    - This raised the Twitter storm alluded to earlier
  - Jager & Leek, *Biostatistics* 2014

Many Labs Replication Project



Klein et al, Social Psychology, 2014

Jager & Leek, 2014

- Text-mining of 77,410 abstracts yielded 5,322 P-values
- Considered a mixture model truncated at 0.05
- Estimated that amongst 'discoveries' 14% are false

$$f(p|a,b,\pi_0) = \pi_0 uniform(0,0.05) + (1-\pi_0) tBeta(a,b,0.05)$$

Estimation using the EM algorithm

But one must be careful

- These studies suggest that a common threshold of 5% seems to be associated with a manageable false positive rate
- This does not mean that the threshold is right
  - It might reflect (say) that most P-values are either >0.05 or <<0.05
  - The situation might be capable of improvement using a different threshold
- Also, are false negatives without cost?

My Conclusion

- P-values *per se* are not the problem
- There may be a harmful culture of 'significance' however this is defined
- P-values have a limited use as rough and ready tools using little structure
- Where you have more structure you can often do better
  - Likelihood, Bayes etc
  - Point estimates and standard errors are extremely useful for future research synthesizers and should be provided regularly