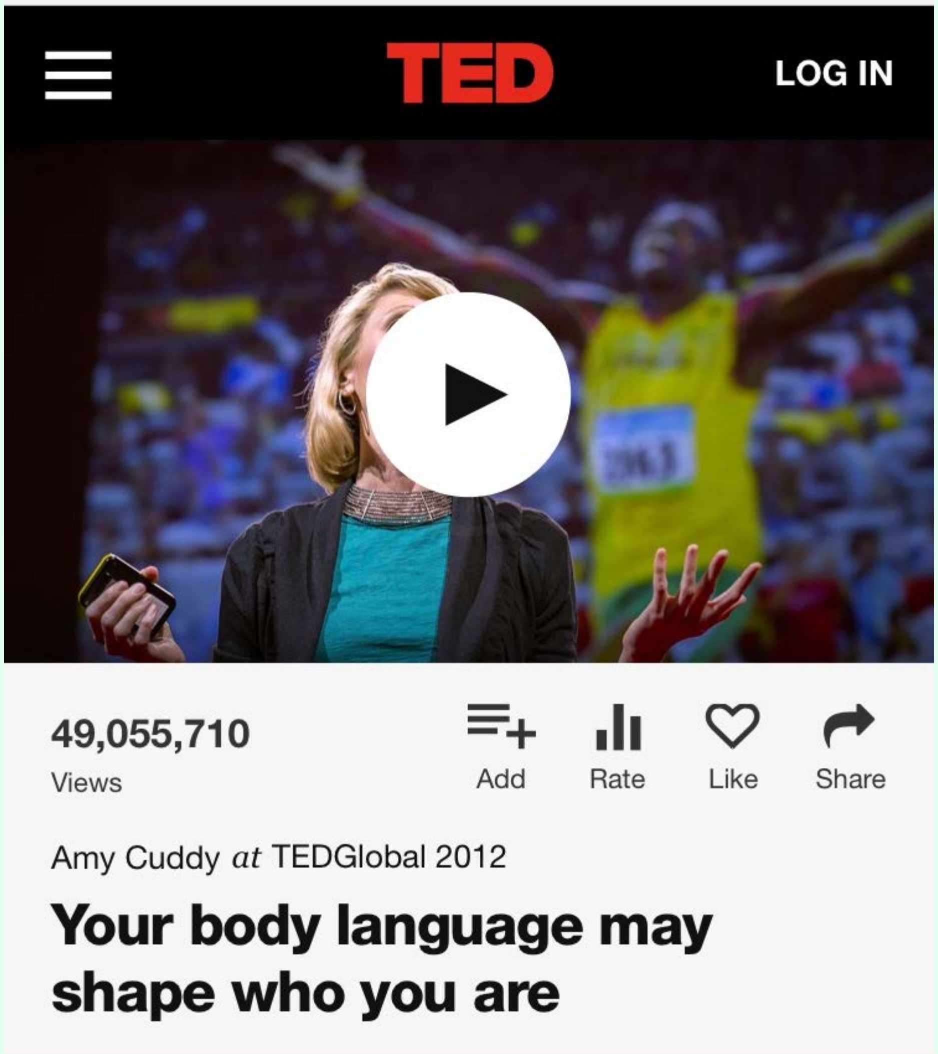
Which results can we trust?
Using replications and prediction markets to assess the reliability of scientific results

SUHF annual conference 2019 at Karolinska Institutet

Anna Dreber Almenberg
Department of Economics
Stockholm School of Economics

Power posing

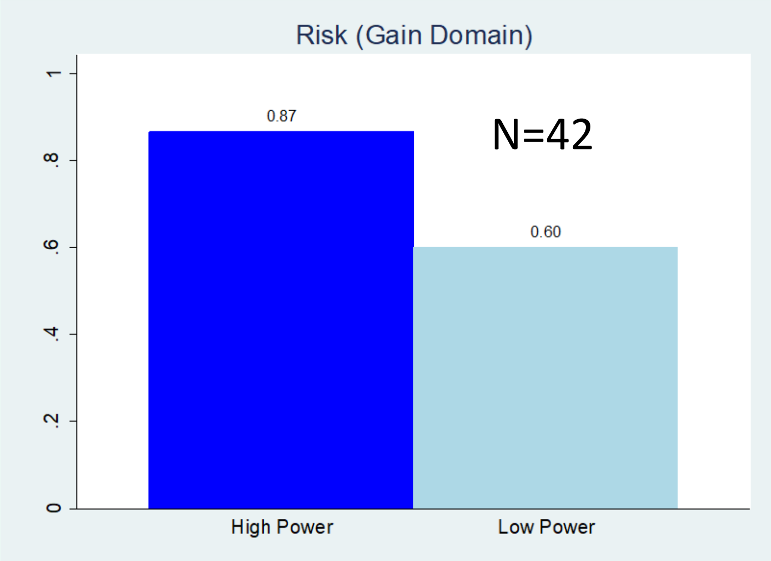


49,055,710
Views

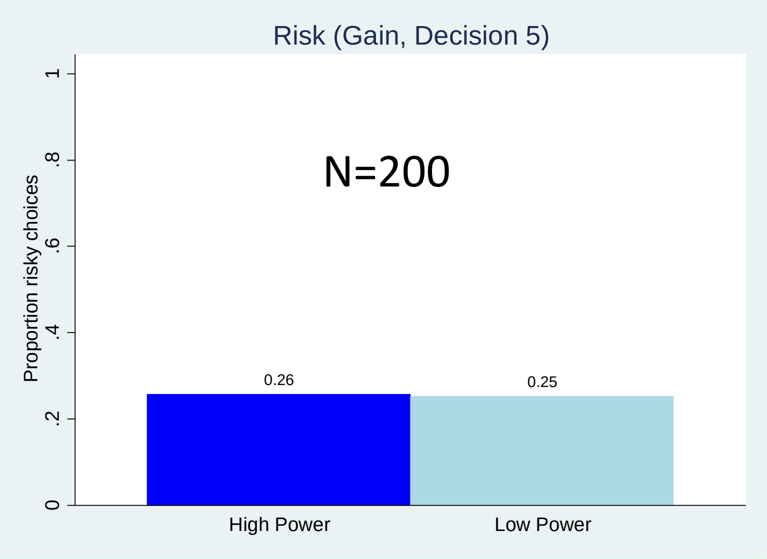
Amy Cuddy *at* TEDGlobal 2012

Your body language may shape who you are

Carney et al. 2010, Ranehill et al. 2015


Risk (Gain Domain)

N=42

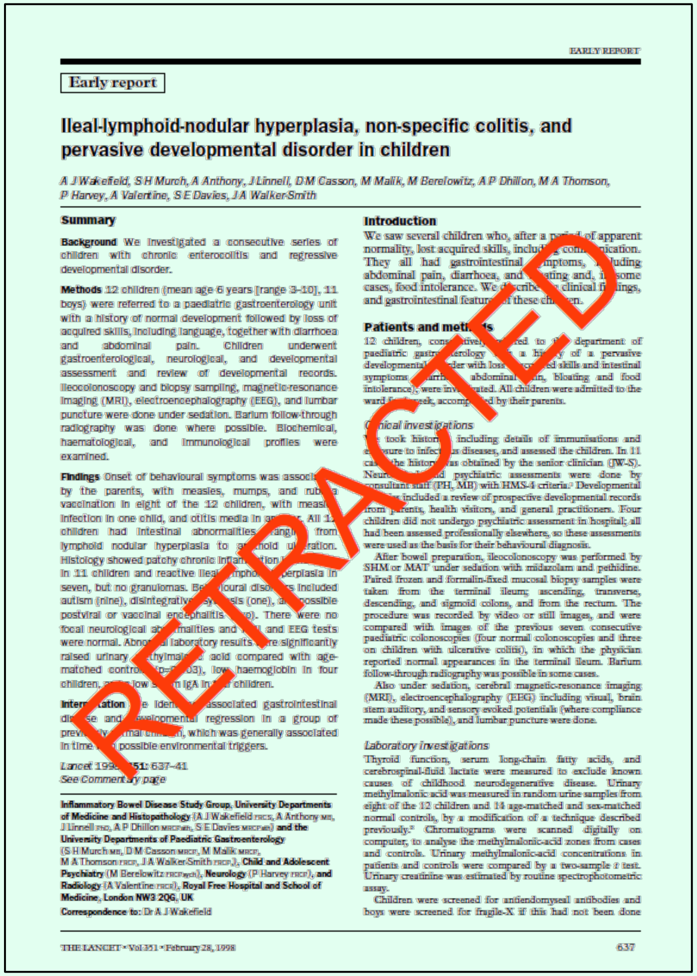

Risk (Gain, Decision 5)

N=200

- How “researcher degrees of freedom” and low statistical power have lead to a replication crisis and how we should design studies and do pre-analysis plans to solve this problem

- Not only an experimental problem
- Not only a social science problem

False results



EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

RETRACTED

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls (p=0·003), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Lancet 1998; **351:** 637–41

See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[1] Developmental histories included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

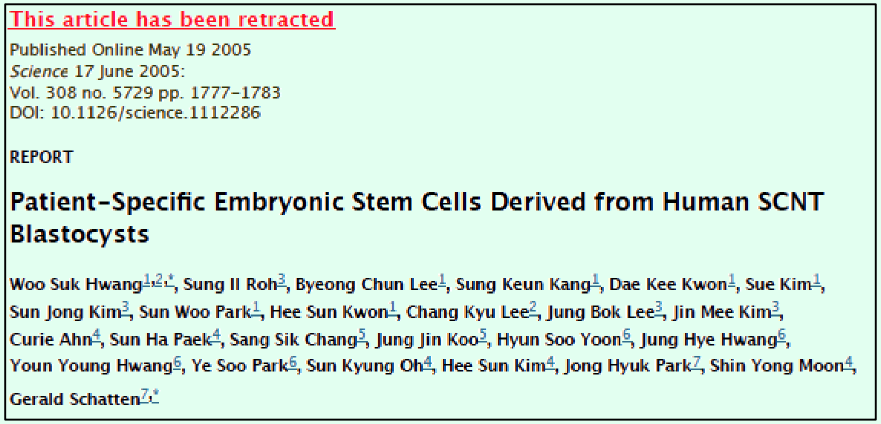
Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[2] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample *t* test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomyseal antibodies and boys were screened for fragile-X if this had not been done

THE LANCET • Vol 351 • February 28, 1998

637



This article has been retracted

Published Online May 19 2005

Science 17 June 2005:

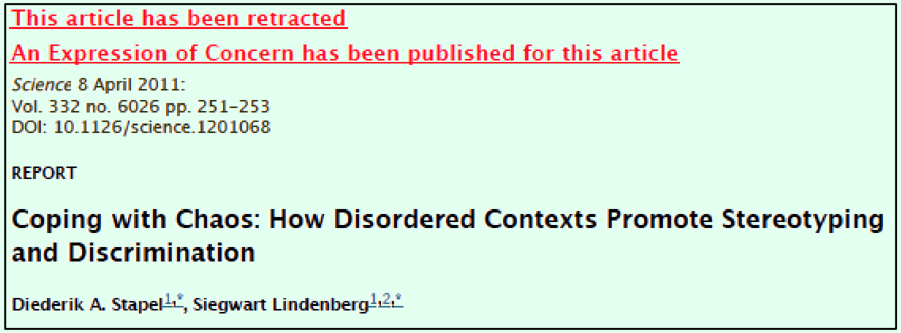
Vol. 308 no. 5729 pp. 1777–1783

DOI: 10.1126/science.1112286

REPORT

Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts

Woo Suk Hwang[1,2,*], Sung Il Roh[3], Byeong Chun Lee[1], Sung Keun Kang[1], Dae Kee Kwon[1], Sue Kim[1], Sun Jong Kim[3], Sun Woo Park[1], Hee Sun Kwon[1], Chang Kyu Lee[2], Jung Bok Lee[3], Jin Mee Kim[3], Curie Ahn[4], Sun Ha Paek[4], Sang Sik Chang[5], Jung Jin Koo[5], Hyun Soo Yoon[5], Jung Hye Hwang[6], Youn Young Hwang[6], Ye Soo Park[6], Sun Kyung Oh[4], Hee Sun Kim[4], Jong Hyuk Park[7], Shin Yong Moon[4], Gerald Schatten[7,*]



This article has been retracted

An Expression of Concern has been published for this article

Science 8 April 2011:

Vol. 332 no. 6026 pp. 251–253

DOI: 10.1126/science.1201068

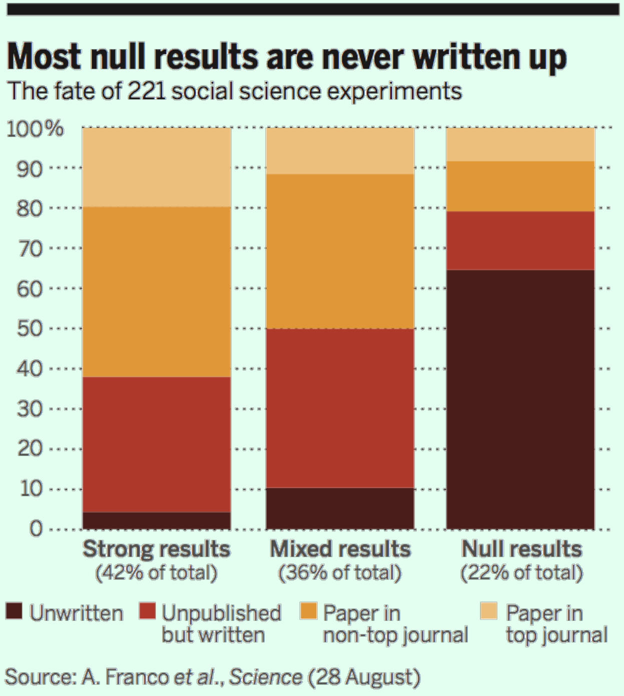
REPORT

Coping with Chaos: How Disordered Contexts Promote Stereotyping and Discrimination

Diederik A. Stapel[1,*], Siegwart Lindenberg[1,2,*]

How many published claims are false?

- False positive results
- False negative results



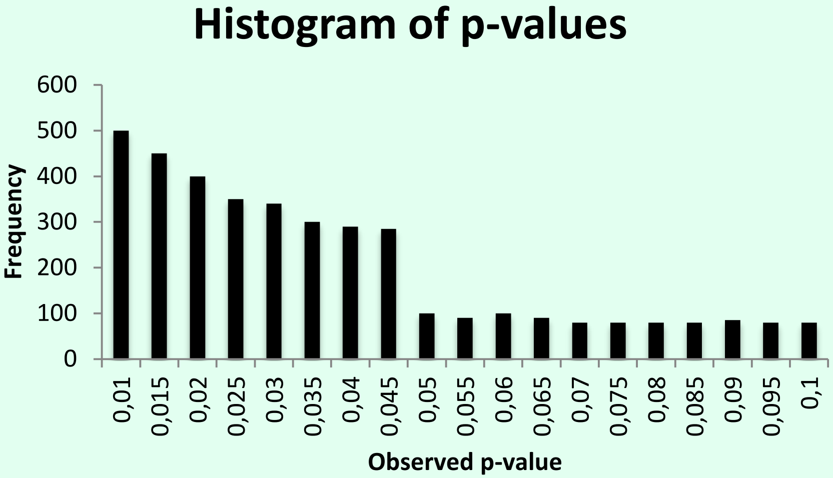
Most null results are never written up

The fate of 221 social science experiments

Source: A. Franco *et al.*, *Science* (28 August)

Ioannidis 2005 PLoS Medicine: Why Most Published Research Findings Are False

“Researcher degrees of freedom”



Ioannidis 2005 Why Most Published Research Findings Are False; Simmons, Nelson and Simonsohn 2011 False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant; Gelman and Loken 2013 The Garden of Forking Paths

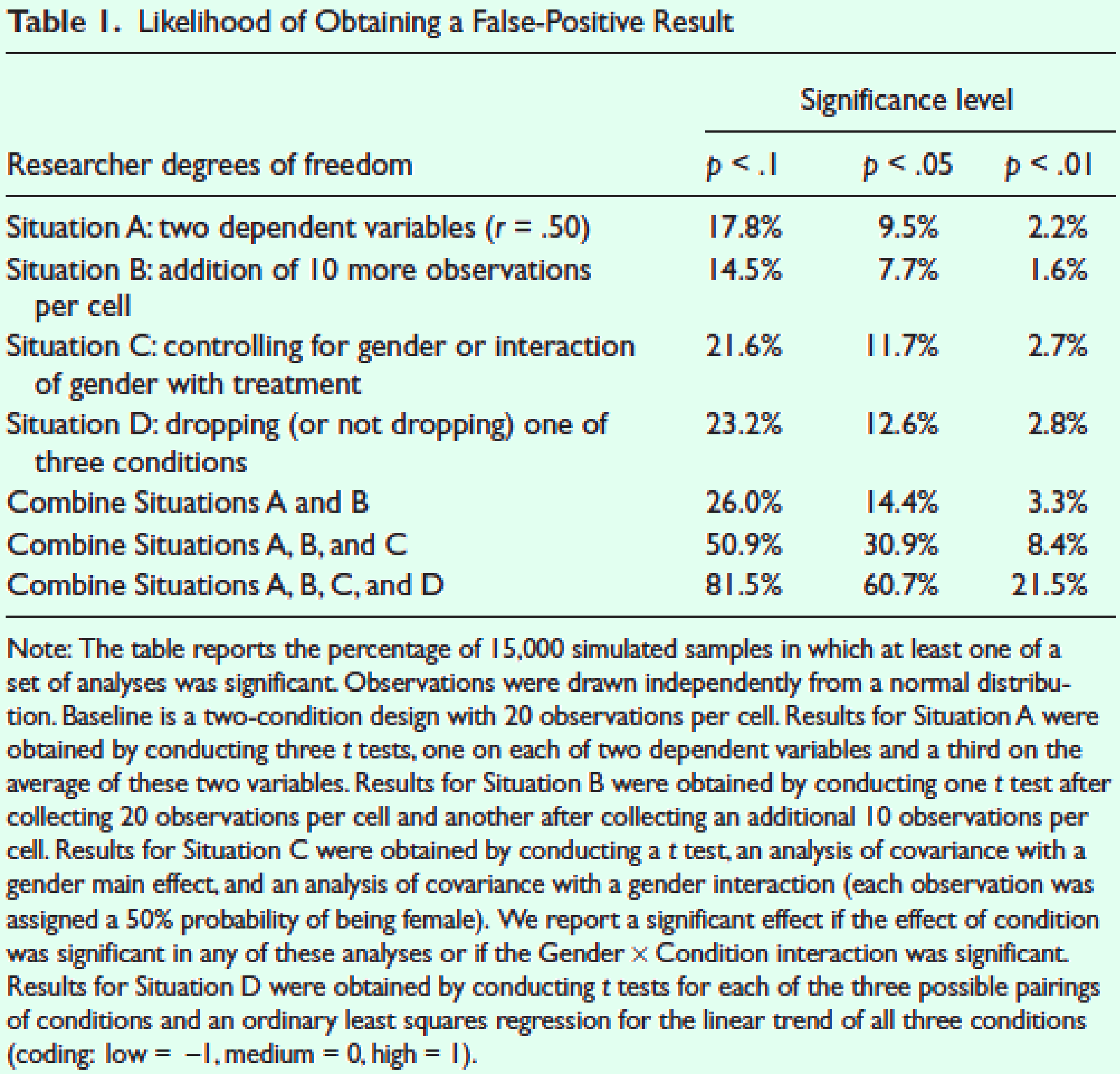


Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables (r = .50)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three *t* tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one *t* test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a *t* test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender × Condition interaction was significant. Results for Situation D were obtained by conducting *t* tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = −1, medium = 0, high = 1).

Simmons, JP, LD Nelson, U Simonsohn, 2011, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22(11): 1359-1366.

Forking

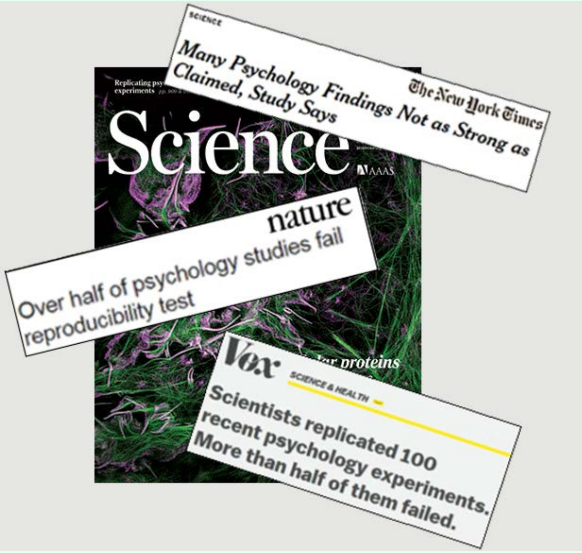
- Multiple testing problem where the universe of tests is not clear
- The data decide the analysis
- Beware subgroup analyses etc
- P-values are meaningless

Which results can we trust?

- Depends on
  - P-values and power
  - Publication bias
  - Researcher degrees of freedom
  - Priors
    - Probability of a hypothesis to be true (“prior”)
    - Typically subjective and unaccessible

How big is the problem?

(In some of the quantitative empirical social sciences)



Science

The New York Times

Many Psychology Findings Not as Strong as Claimed, Study Says

nature

Over half of psychology studies fail reproducibility test

Vox SCIENCE & HEALTH

Scientists replicated 100 recent psychology experiments. More than half of them failed.

A minimal synthetic genome pp. 1360 & 1414

Unearthing an epic battle in Bronze Age Germany p. 1364

Ad hoc international ethics review p. 1369

Science

$15 25 MARCH 2016 sciencemag.org

AAAS

The Economist World politics Business & finance Economics Science

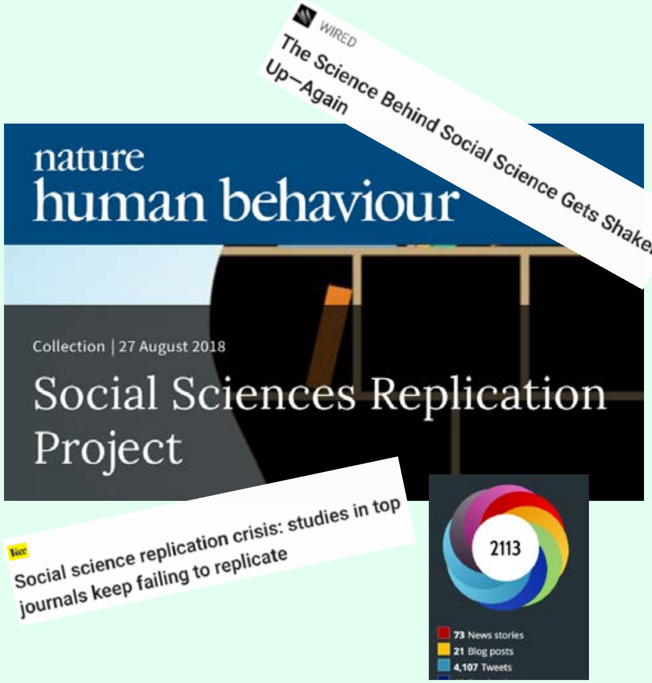
Economics

A far from dismal outcome

Microeconomists' claims to be doing real science turn out to be true

Recycling light

Perovskite solar cells reuse emitted photons pp. 1401 & 1430



WIRED

The Science Behind Social Science Gets Shake Up—Again

nature human behaviour

Collection | 27 August 2018

Social Sciences Replication Project

Vox

Social science replication crisis: studies in top journals keep failing to replicate

2113

- 🟥 73 News stories
- 🟨 21 Blog posts
- 🟦 4,107 Tweets

Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." *Science*.
Camerer et al. (2016) "Evaluating replicability of laboratory experiments in economics." *Science.*
Camerer et al. (2018) "Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015." *Nature Human Behaviour.*

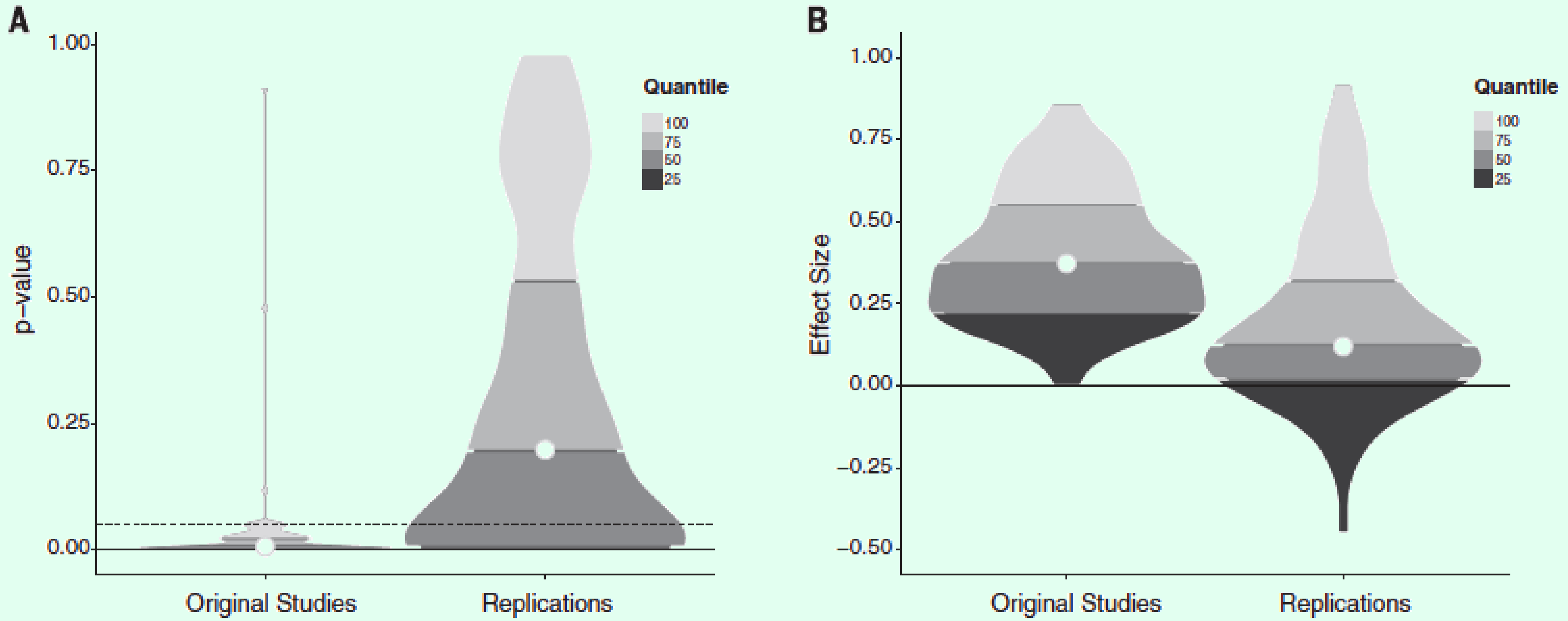
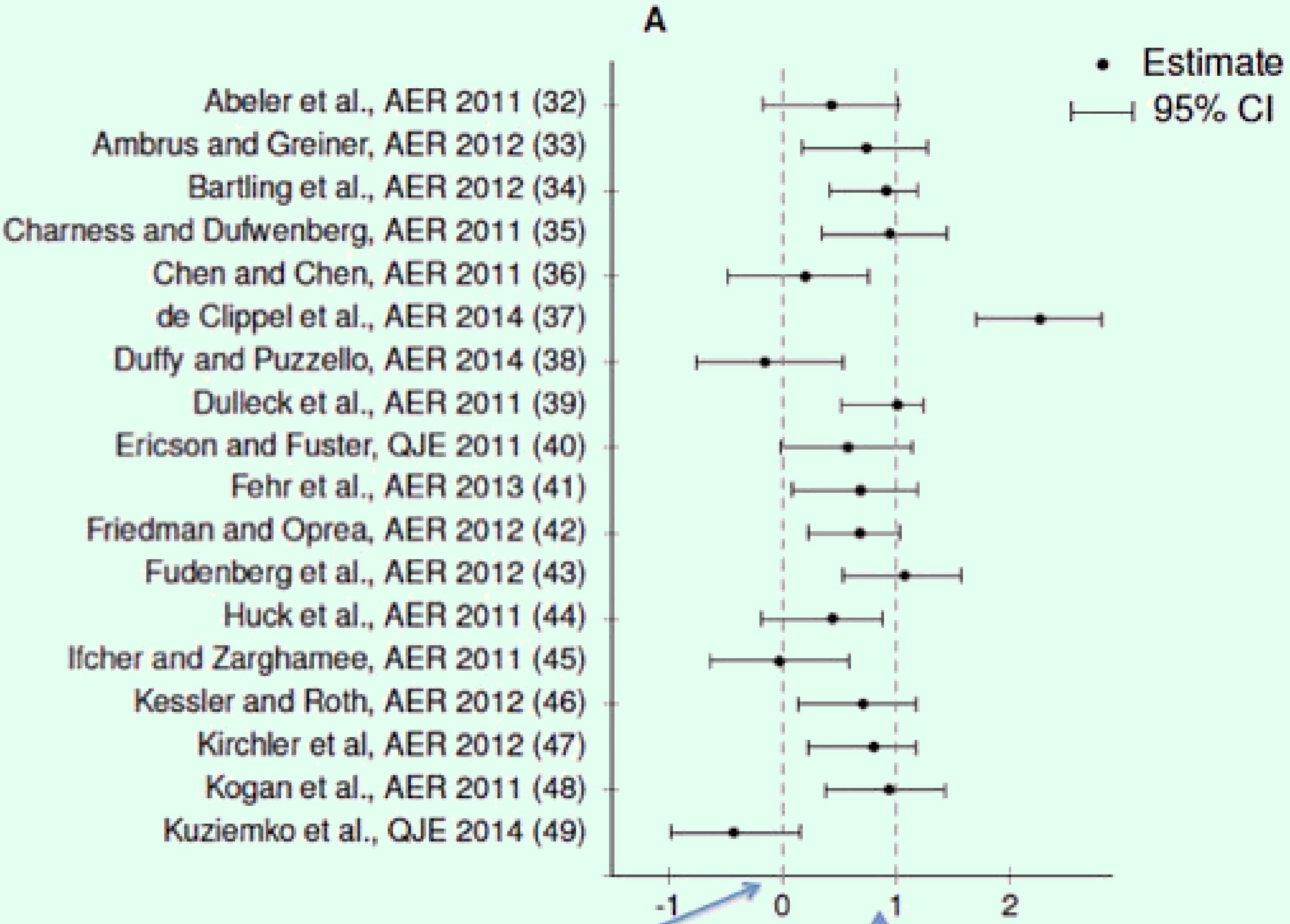


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

35/97 positive results replicate
Relative effect size about 50%

Open Science Collaboration (2015). “Estimating the Reproducibility of Psychological Science.” *Science*, 349(6251).

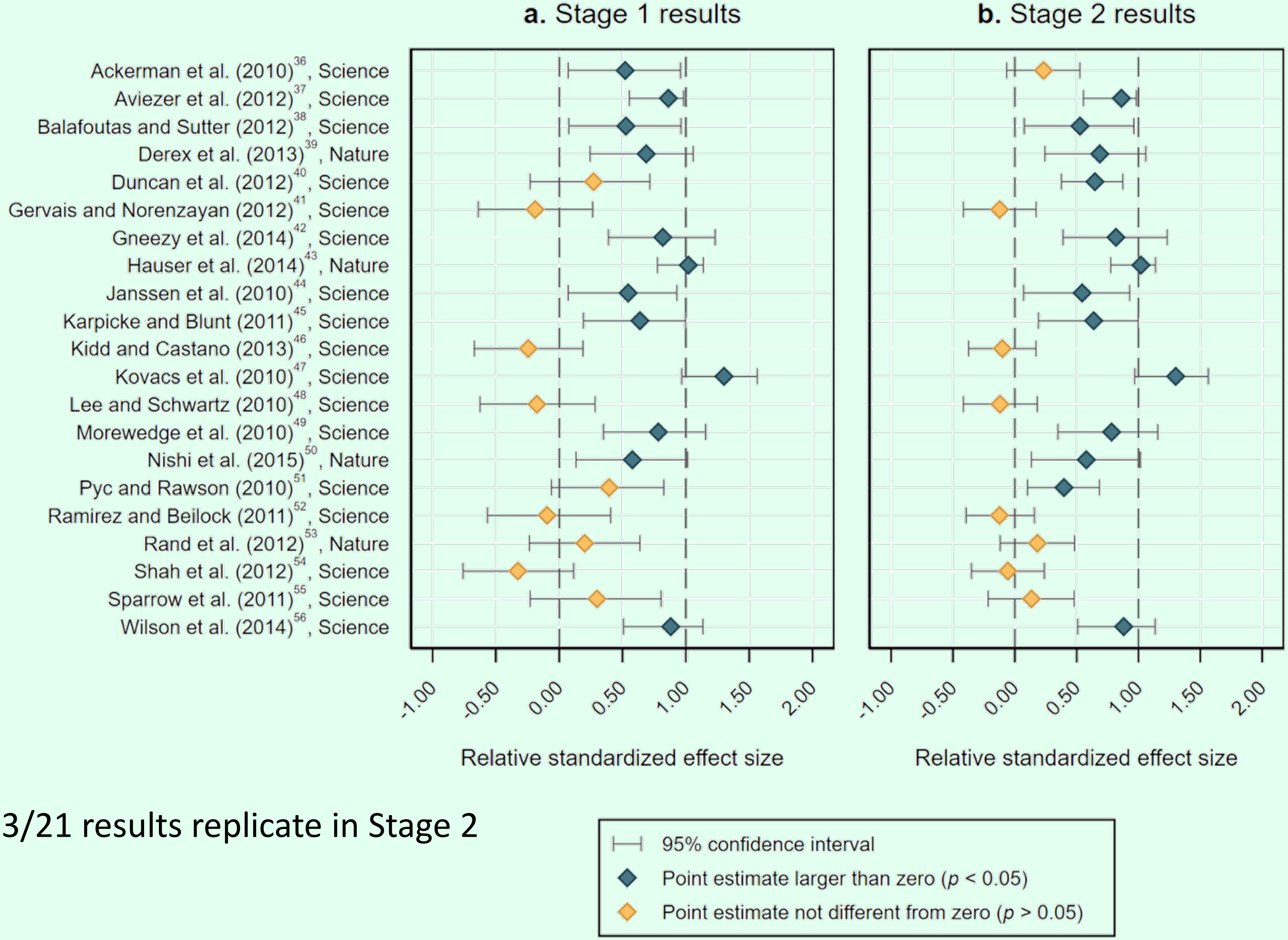


0= no effect

1=same effect as in original study

11/18 results replicate
Relative effect size about 60%

Camerer et al. 2016 *Science*

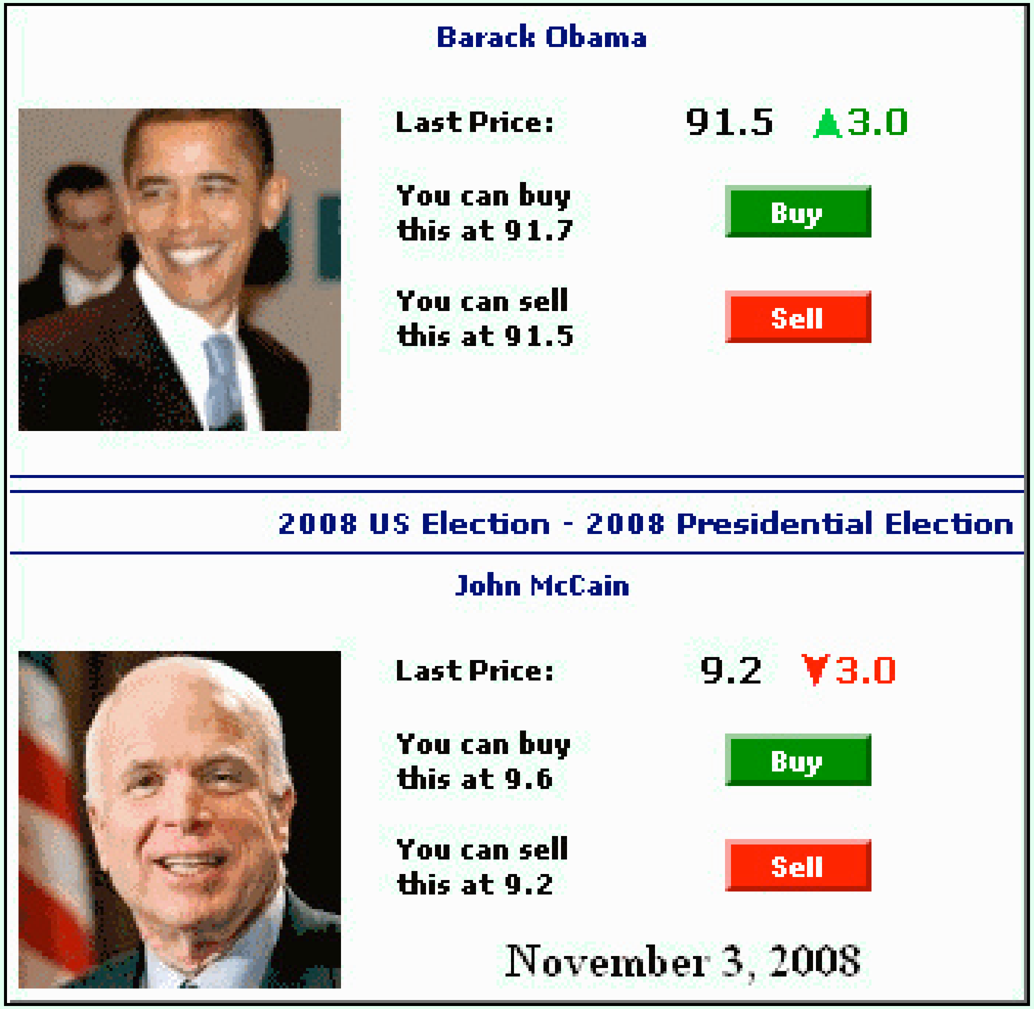


13/21 results replicate in Stage 2

Mean relative effect size: 50%. For 13 studies that replicated: 74%, for the rest, 0%

Camerer et al. 2018 *Nature Human Behaviour*

“Could gambling save science?”



Barack Obama

Last Price: 91.5 ▲3.0

You can buy this at 91.7 Buy

You can sell this at 91.5 Sell

2008 US Election - 2008 Presidential Election

John McCain

Last Price: 9.2 ▼3.0

You can buy this at 9.6 Buy

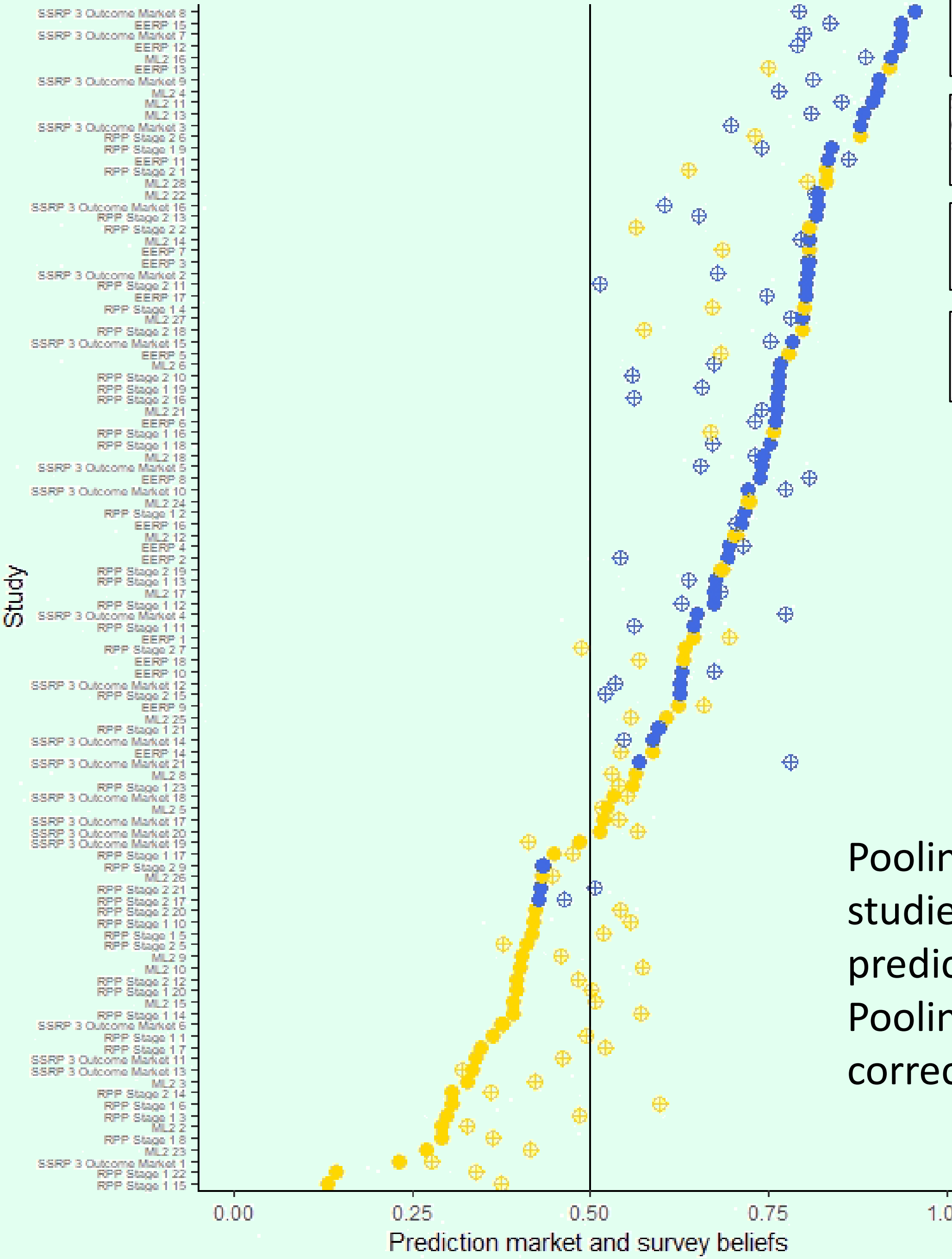
You can sell this at 9.2 Sell

November 3, 2008

Hanson 1995 *Social Epistemology*

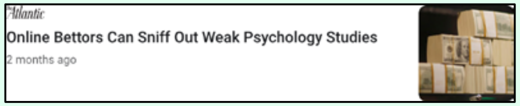
Our prediction markets on replications

- 10 days – 2 weeks
- USD 50-100
- 50-100 participants
- Central hypothesis
- Binary outcomes
- Price: predicted probability of the outcome occuring
- Participants get replication reports
- Also survey questions



The power of prediction markets

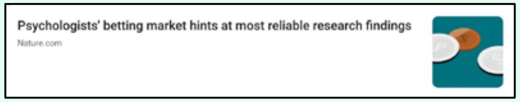
Nature.com



Atlantic

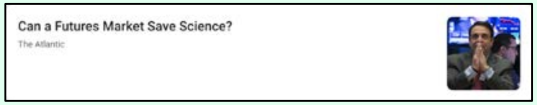
Online Bettors Can Sniff Out Weak Psychology Studies

2 months ago



Psychologists' betting market hints at most reliable research findings

Nature.com



Can a Futures Market Save Science?

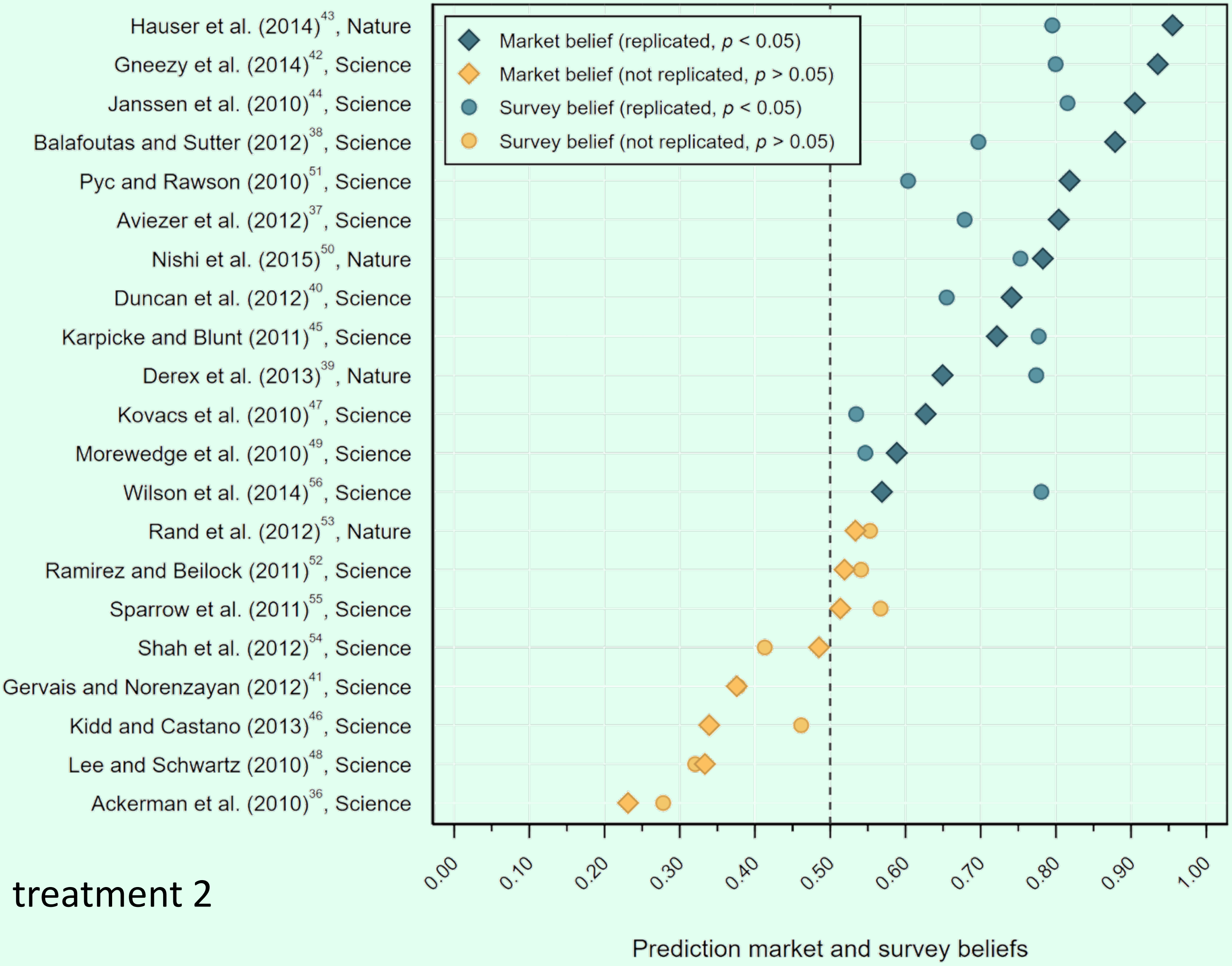
The Atlantic

Pooling 4 prediction market studies: 73% (76/104) correct prediction rate

Pooling 4 surveys: 66% (68/103) correct prediction rate

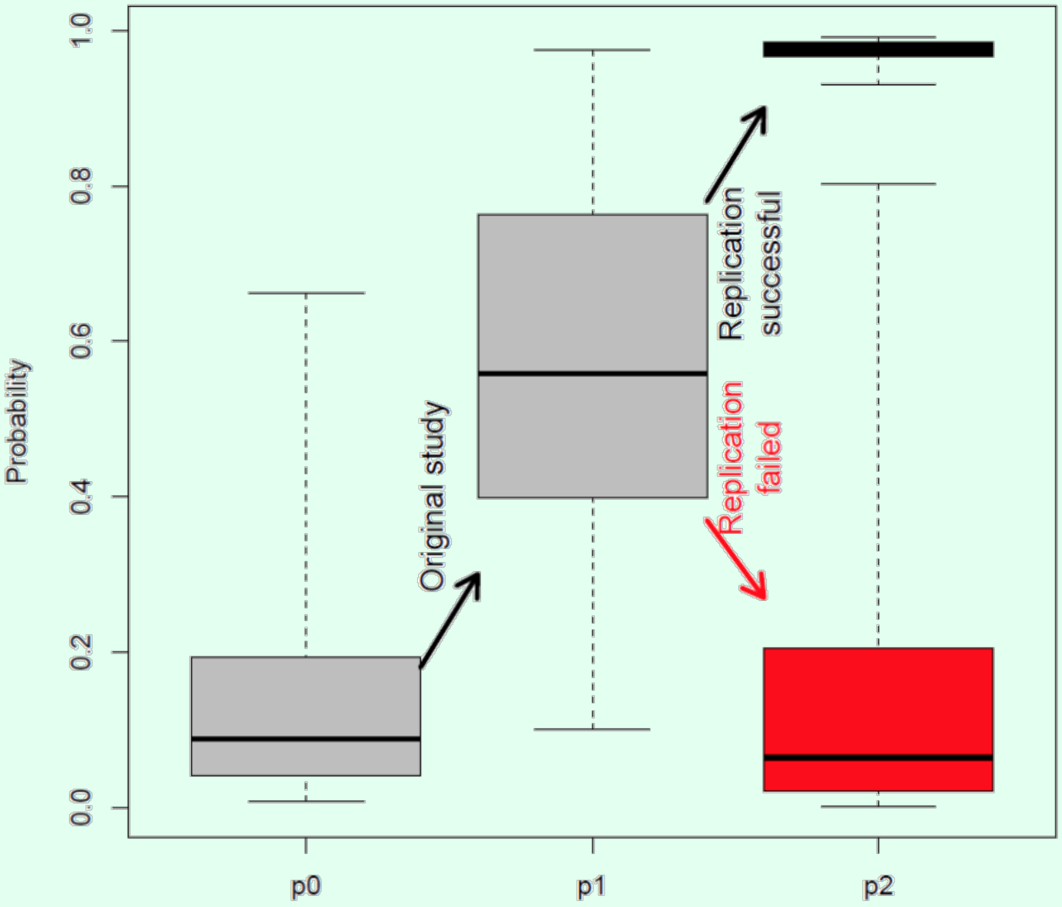
Work in progress

Prediction markets results *Nature* and *Science*



From treatment 2

Probability of hypothesis being true at 3 stages of testing for RPP



- Initial priors are low (median 8.8%)
- Positive result in initial publication moves prior to intermediate level (median 56%)
- If successful replication, probability moves up (median 98%)
- If failed replication, probabiliby close to initial prior (median 6.3%)

Whiskers: range
Boxes: 1st to 3rd quartiles
Thick lines: medians

Dreber et al. 2015 PNAS

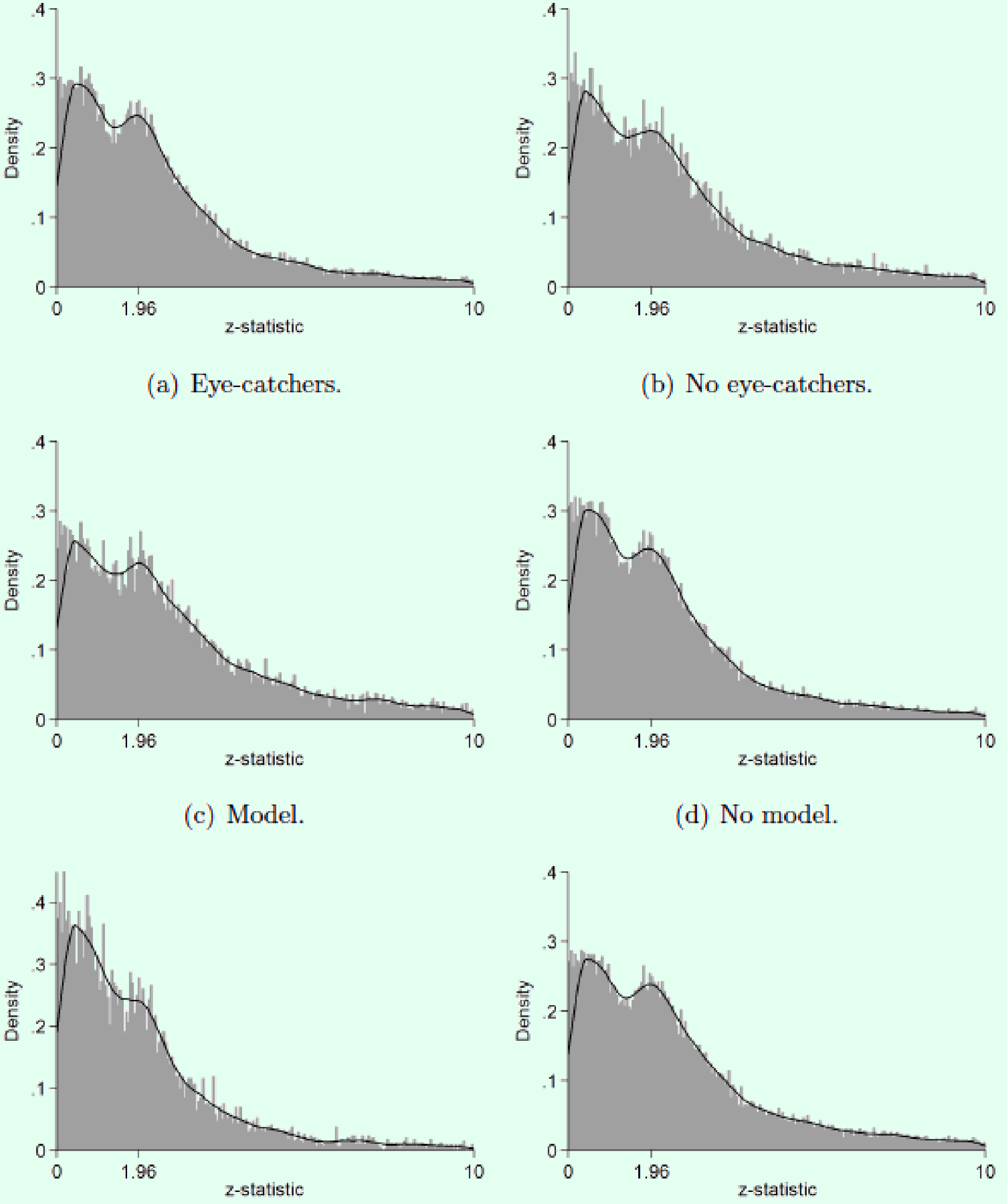
What have we learned?

- Common false interpretation of p<0.05: 95% probability of hypothesis being true
- For this to be the case, a p<0.05 finding needs to supported in a high-powered replication
- Meta-analyses will also have inflated effect sizes – we need replications
- Are the incentives for replications appropriate?
- There is something systematic about results that fail to replicate – and experts “know” this
  - So why are so many false results published?

See, e.g., 2019 book chapter by Camerer, Dreber and Johannesson for more

Other thoughts

- Pre-analysis plans
- Problems probably worse for non-experimental work
- Higher power and team science
  - Munafo et al. 2017 Nature Human Behaviour
- $p<0.005$



(a) Eye-catchers.

(b) No eye-catchers.

(c) Model.

(d) No model.

(e) Lab. experiments or RCT data.

(f) Other data.

Brodeur et al 2016

Sources: AER, JPE, and QJE (2005-2011). Distributions are unweighted and plotted using de-rounded statistics. Lines correspond to kernel density estimates.

p<0.005

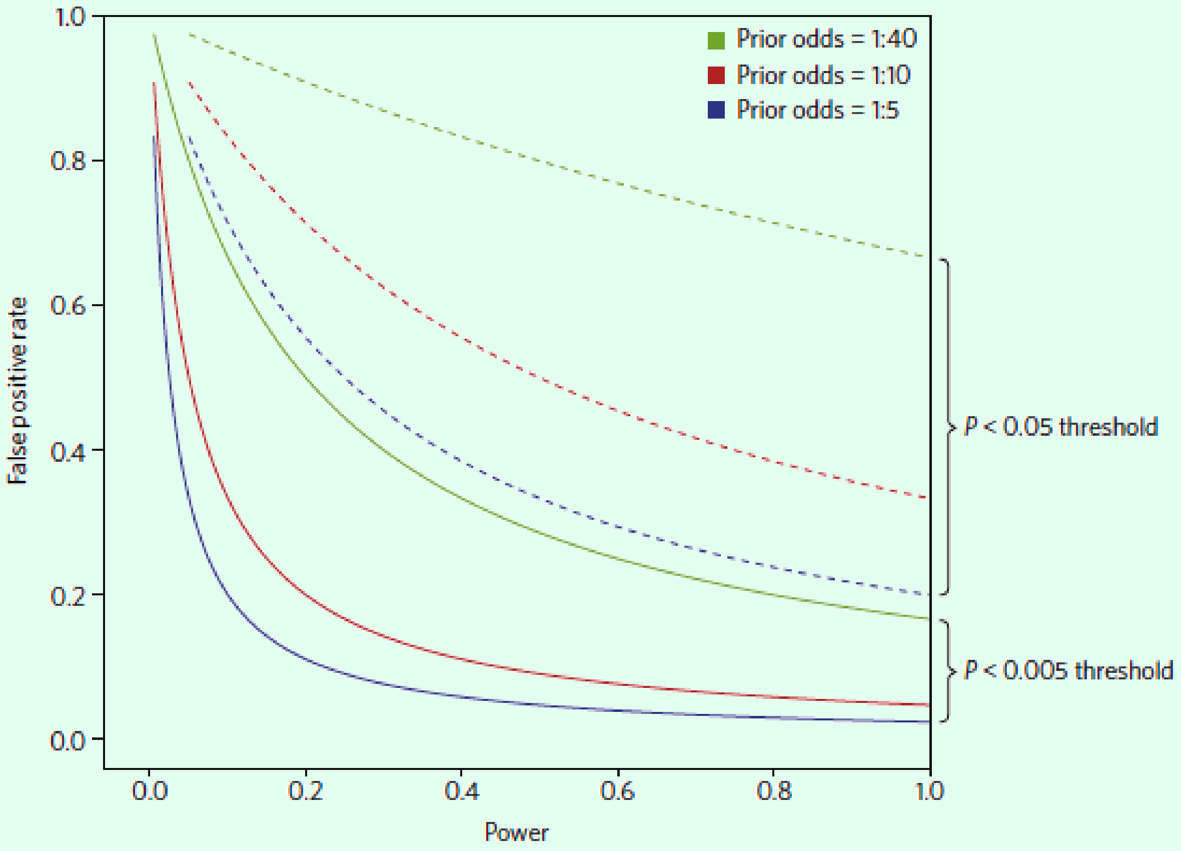


Fig. 2 | Relationship between the *P* value threshold, power, and the false positive rate. Calculated according to equation (2), with prior odds defined as $\frac{1-\phi}{\phi} = \frac{\Pr(H_1)}{\Pr(H_0)}$. For more details, see the Supplementary Information.

Benjamin et al. 2018 “Redefine Statistical Significance” Nature Human Behavior

Thanks!

anna.dreber@hhs.se

www.replicationmarkets.com