WELCOME TO DAY 1

Amy Perfors

SCHOOL OVERVIEW

Entire process of performing and analysing an experiment from beginning to end

Day 0: R Bootcamp

Day 1: Experiment

Day 2: Data analysis in R

Day 3: Modelling

Day 4: Experience sampling

Day 5: Wrap-up

DAY 1: EXPERIMENT

Not just coding it up, but all workflow stuff up to running it

1. Background: replicability and proper procedure
2. Workflow and organisation
3. Experiment design
4. Coding experiment
5. Ethics and pre-registration
6. Hosting experiment on a server
7. Downloading data

PLAN FOR TODAY

- The replication crisis (a history)
- What's going on?
- Possible solutions

To anonymously make questions or comments, go to www.menti.com and use the code 36 43 0

2011 WAS A DIFFICULT YEAR

ESP WAS PROVEN TO EXIST

Experiment 1: N=100

Predicted erotic pictures significantly more frequently than the 50% hit rate expected by chance (53.1%, t(99)=2.51, p=0.01, d=0.25).

Experiment 2: N=150, negative only

Also different from chance (51.7%, t(149)=2.39, p=0.009, d=0.2).

Which location has an picture behind it?

(people predicted in advance)

Some of the pictures were erotic, some were negative, some were neutral

Everything completely randomised

Bem et al (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect

ESP WAS PROVEN TO EXIST

Experiment 1: N=100

Predicted erotic pictures significantly more frequently than the 50% hit rate expected by chance (53.1%, t(99)=2.51, p=0.01, d=0.25).

Experiment 2: N=150, negative only

Also different from chance (51.7%, t(149)=2.39, p=0.009, d=0.2).

Experiment 3: N=100, retroactive priming

People were 15.0ms faster on congruent trials (t(96)=2.55, p=0.006, d=0.25)

Experiment 4: N=100, retroactive priming

People were 16.5ms faster on congruent trials (t(98)=2.03, p=0.023, d=0.2)

Experiment 5: N=100, retroactive habituation

People preferred the target more (t(99)=2.23, p=0.014, d=0.22)

Experiment 6: N=150, retroactive habituation

People preferred the target more with negative (51.8%, t(149)=1.80, p=0.037, d=0.15). Less with erotic pairs (48.2%, t(149)=1.77, p=0.039, d=0.14)

Experiment 7: N=200, retroactive boredom

Not significant (49.1%, t(199)=1.31, p=0.096, d=0.09)

Experiment 8: N=100, recall facilitation

Recall facilitated by showing it in the future (t(99)=1.92, p=0.029, d=0.19)

Experiment 9: N=50, recall facilitation

Recall facilitated by showing it in the future (t(49)=2.96, p=0.002, d=0.42)

Bem et al (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect

LUCKILY (?) THIS DID NOT REPLICATE

OPEN ACCESS Freely available online

PLoS one

Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect

Stuart J. Ritchie[1]*, Richard Wiseman[2], Christopher C. French[3]

1 Psychology Department, The University of Edinburgh, Edinburgh, United Kingdom, **2** School of Psychology, University of Hertfordshire, Hatfield, United Kingdom, **3** Anomalistic Psychology Research Unit, Goldsmiths, University of London, London, United Kingdom

Abstract

Nine recently reported parapsychological experiments appear to support the existence of precognition. We describe three pre-registered independent attempts to exactly replicate one of these experiments, 'retroactive facilitation of recall', which examines whether performance on a memory test can be influenced by a post-test exercise. All three replication attempts failed to produce significant effects (combined $n = 150$; combined $p = .83$, one-tailed) and thus do not support the existence of psychic ability.

Journal of Personality and Social Psychology
2012, Vol. 103, No. 6, 933–948

0022-3514/12/$12.00 DOI: 10.1037/a0029709

Correcting the Past: Failures to Replicate Psi

Jeff Galak
Carnegie Mellon University

Robyn A. LeBoeuf
University of Florida

Leif D. Nelson
University of California, Berkeley

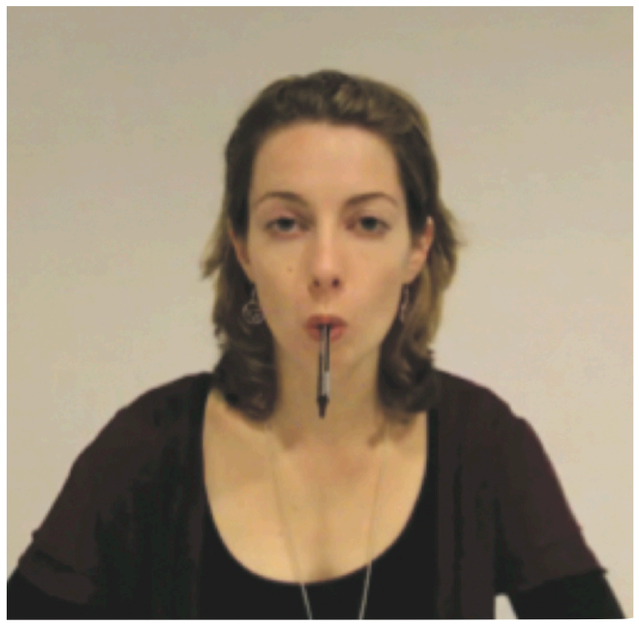
Joseph P. Simmons
University of Pennsylvania

Across 7 experiments ($N = 3,289$), we replicate the procedure of Experiments 8 and 9 from Bem (2011), which had originally demonstrated retroactive facilitation of recall. We failed to replicate that finding. We further conduct a meta-analysis of all replication attempts of these experiments and find that the average effect size ($d = 0.04$) is no different from 0. We discuss some reasons for differences between the results in this article and those presented in Bem (2011).

…BUT MANY THINGS DON'T REPLICATE

Smiling makes you happier

(cited 1788 times)



J Pers Soc Psychol. 1988 May;54(5):768-77.

Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis.

Strack F[1], Martin LL, Stepper S.

Author information

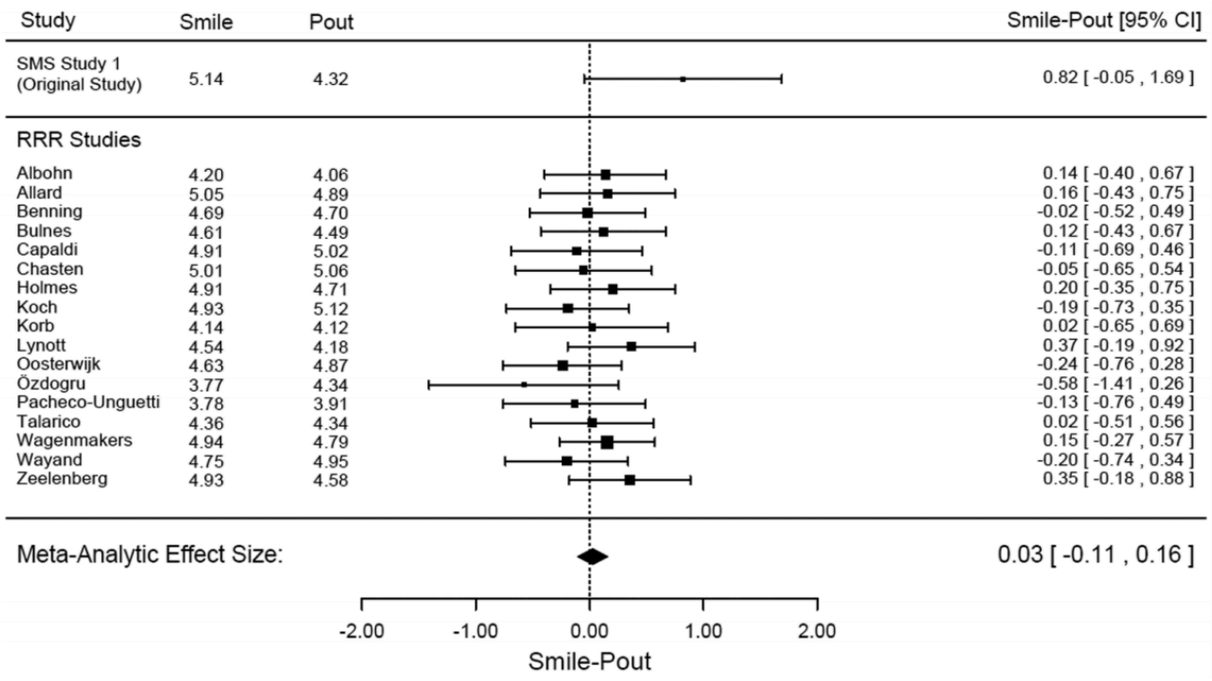
Abstract

We investigated the hypothesis that people's facial activity influences their affective responses. Two studies were designed to both eliminate methodological problems of earlier experiments and clarify theoretical ambiguities. This was achieved by having subjects hold a pen in their mouth in ways that either inhibited or facilitated the muscles typically associated with smiling without requiring subjects to pose in a smiling face. Study 1's results demonstrated the effectiveness of the procedure. Subjects reported more intense humor responses when cartoons were presented under facilitating conditions than under inhibiting conditions that precluded labeling of the facial expression in emotion categories. Study 2 served to further validate the methodology and to answer additional theoretical questions. The results replicated Study 1's findings and also showed that facial feedback operates on the affective but not on the cognitive component of the humor response. Finally, the results suggested that both inhibitory and facilitatory mechanisms may have contributed to the observed affective responses.

…BUT MANY THINGS DON'T REPLICATE

Smiling **doesn't make** you happier

17 independent direct replications, 1894 participants



Perspectives on Psychological Science
2016, Vol. 11(6) 917–928

DOI: 10.1177/1745691616674458
pps.sagepub.com
SAGE

Registered Replication Report: Strack, Martin, & Stepper (1988)

E.-J. Wagenmakers*, T. Beek*, L. Dijkhoff*, Q. F. Gronau,* A. Acosta, R. B. Adams, Jr., D. N. Albohn, E. S. Allard, S. D. Benning, E.-M. Blouin-Hudon, L. C. Bulnes, T. L. Caldwell, R. J. Calin-Jageman, C. A. Capaldi, N. S. Carfagno, K. T. Chasten, A. Cleeremans, L. Connell, J. M. DeCicco, K. Dijkstra, A. H. Fischer, F. Foroni, U. Hess, K. J. Holmes, J. L. H. Jones, O. Klein, C. Koch, S. Korb, P. Lewinski, J. D. Liao, S. Lund, J. Lupianez, D. Lynott, C. N. Nance, S. Oosterwijk, A. A. Ozdoğru, A. P. Pacheco-Unguetti, B. Pearson, C. Powis, S. Riding, T.-A. Roberts, R. I. Rumiati, M. Senden, N. B. Shea-Shumsky, K. Sobocko, J. A. Soto, T. G. Steiner, J. M. Talarico, Z. M. van Allen, M. Vandekerckhove, B. Wainwright, J. F. Wayand, R. Zeelenberg, E. E. Zetzer, and R. A. Zwaan
*Proposing authors

Protocol vetted by: Ursula Hess

Protocol edited by: Daniel J. Simons

Multilab direct replication of: Study 1 from Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768-777.

Data and registered protocols: https://osf.io/pkd65/

Citation: Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., . . . Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*, 917–928.

Abstract

According to the *facial feedback hypothesis*, people's affective responses can be influenced by their own facial expression (e.g., smiling, pouting), even when their expression did not result from their emotional experiences. For example, Strack, Martin, and Stepper (1988) instructed participants to rate the funniness of cartoons using a pen that they held in their mouth. In line with the facial feedback hypothesis, when participants held the pen with their teeth (inducing a "smile"), they rated the cartoons as funnier than when they held the pen with their lips (inducing a "pout"). This seminal study of the facial feedback hypothesis has not been replicated directly. This Registered Replication Report describes the results of 17 independent direct replications of Study 1 from Strack et al. (1988), all of which followed the same vetted protocol. A meta-analysis of these studies examined the difference in funniness ratings between the "smile" and "pout" conditions. The original Strack et al. (1988) study reported a rating difference of 0.82 units on a 10-point Likert scale. Our meta-analysis revealed a rating difference of 0.03 units with a 95% confidence interval ranging from −0.11 to 0.16.

Table 2. Bayes Factors for Each of the 17 Replication Studies

Replication lab	Default BF_{10}	Replication BF_{r0}
Albohn	0.281	0.297
Allard	0.300	0.329
Benning	0.189	0.190
Bulnes	0.300	0.343
Capaldi	0.150	0.149
Chasten	0.191	0.199
Holmes	0.401	0.499
Koch	0.134	0.139
Korb	0.219	0.232
Lynott	0.713	0.993
Oosterwijk	0.115	0.121
Özdoğru	0.106	0.124
Pacheco-Unguetti	0.146	0.144
Talarico	0.215	0.222
Wagenmakers	0.356	0.406
Wayand	0.126	0.129
Zeelenberg	0.773	1.136

…BUT MANY THINGS DON'T REPLICATE

A pair of eyes makes people more honest

(cited 1049 times)

VS

BIOLOGY LETTERS

search

Home Content Information for About us Sign up Submit

Cues of being watched enhance cooperation in a real-world setting

Melissa Bateson, Daniel Nettle, Gilbert Roberts

Published 22 September 2006. DOI: 10.1098/rsbl.2006.0509

…BUT MANY THINGS DON'T REPLICATE

A pair of eyes
doesn't make people
more honest

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

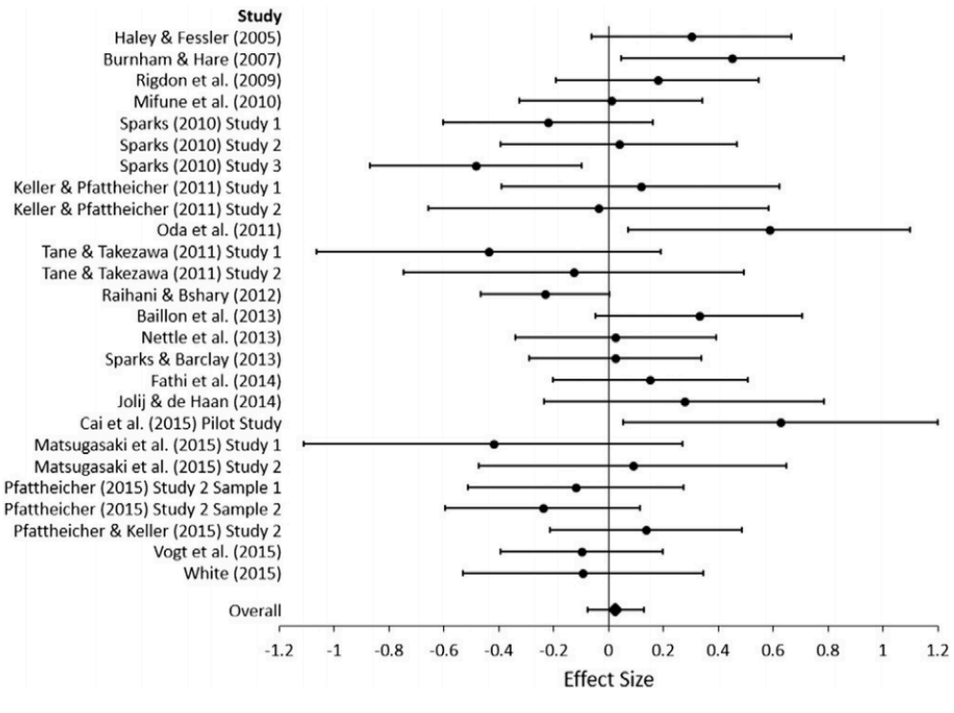
Review Article

Artificial surveillance cues do not increase generosity: two meta-analyses

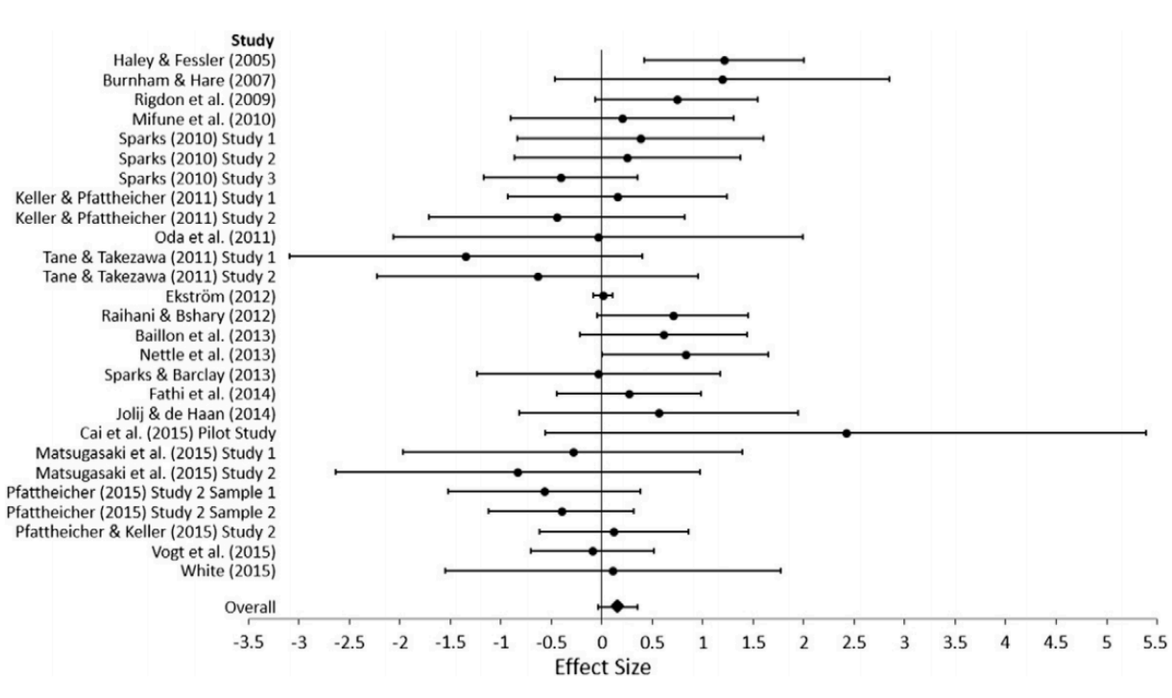
Stefanie B. Northover [a,b,*], William C. Pedersen [c], Adam B. Cohen [b], Paul W. Andrews [a]

[a] Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada

Meta-analysis, 26 experiments, 2700 people

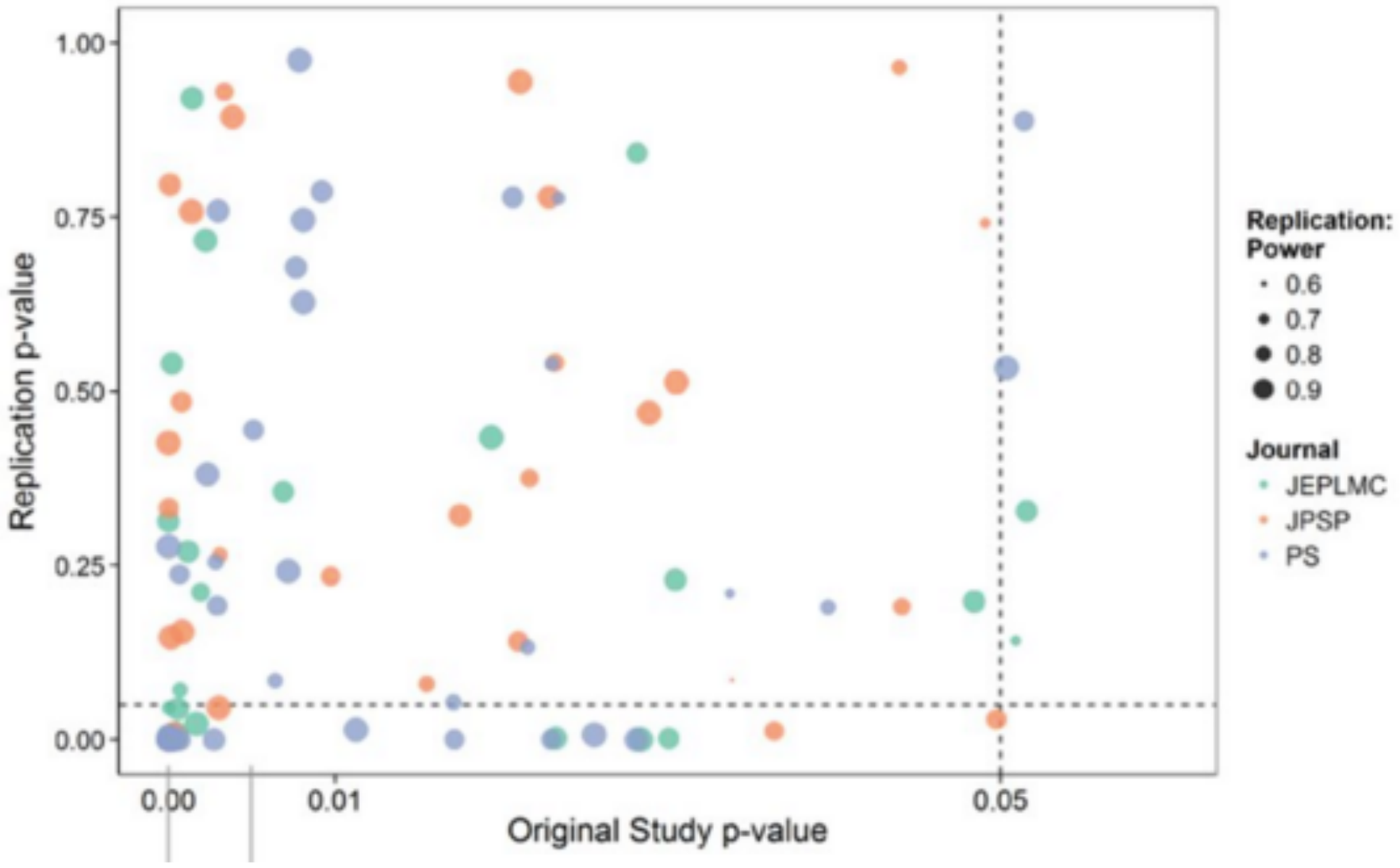


Meta-analysis, 27 experiments, nearly 20,000 people



THIS IS A REAL PROBLEM

Conducted replications of 100 original studies using the same materials and procedure (as much as possible)

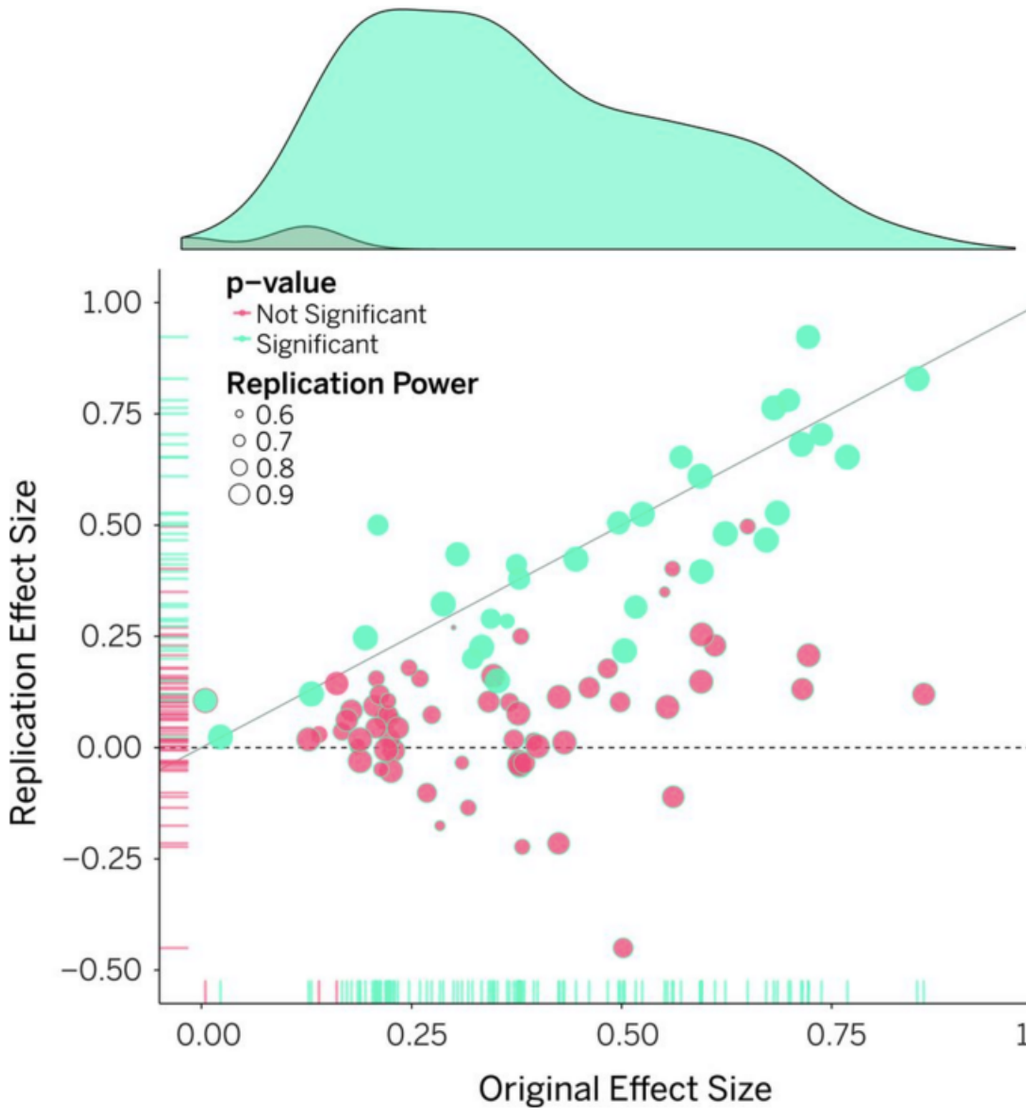


Many originally-significant findings didn't replicate

Nosek et al., (2015). Estimating the reproducibility of psychological science. Science

THIS IS A REAL PROBLEM

Conducted replications of 100 original studies using the same materials and procedure (as much as possible)

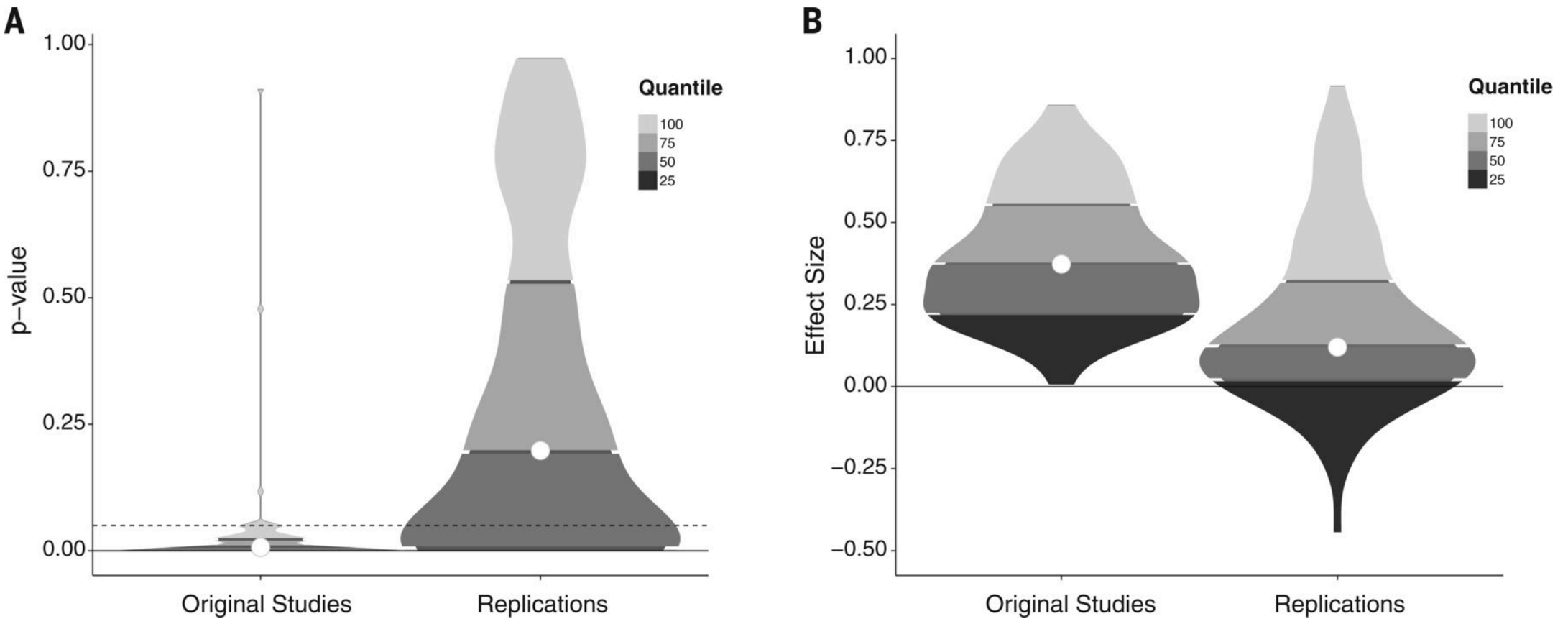


It extends to effect size

Nosek et al., (2015). Estimating the reproducibility of psychological science. Science

THIS IS A REAL PROBLEM

Conducted replications of 100 original studies using the same materials and procedure (as much as possible)



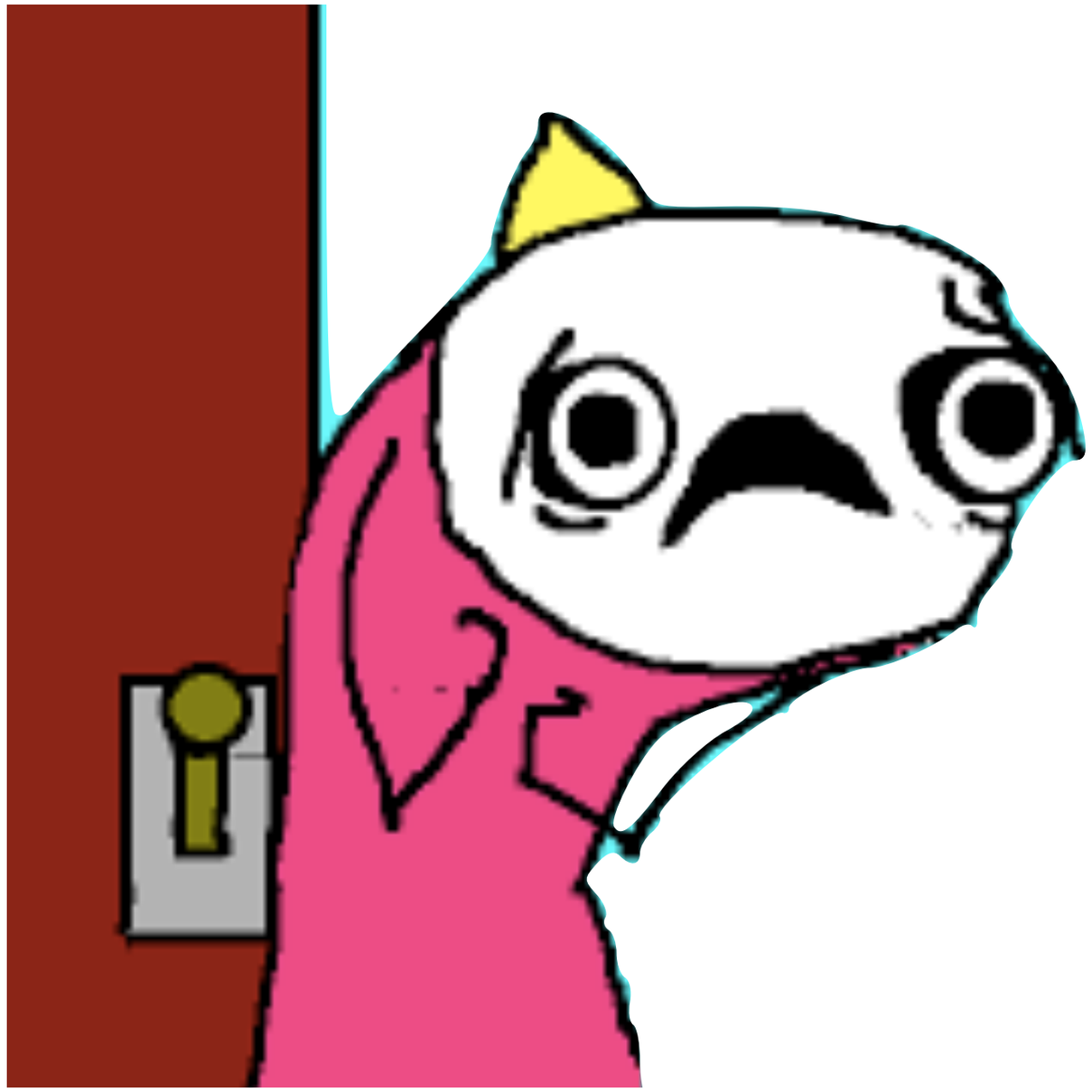
Nosek et al., (2015). Estimating the reproducibility of psychological science. Science

THIS IS A REAL PROBLEM

It's not just psychology…

- Begley & Ellis, 2012: Nature. Researchers tried to reproduce 53 landmark cancer findings: succeeded in only 6 (i.e., 11%)
- Ioannidis, 2005. Looked at 49 highly cited (>1000 times) research studies. 45 claimed intervention was effective. 16% contradicted by later studies, 16% found stronger effects than later studies, 44% replicated, 24% unchallenged.

WHAT'S GOING ON??



SOME POSSIBILITIES

- Reflective of the real-world
- Poor incentive structures
- Poor statistical practice

REAL WORLD?

Maybe different people just behave differently in different contexts

Lower replication in social psychology than cognitive psychology

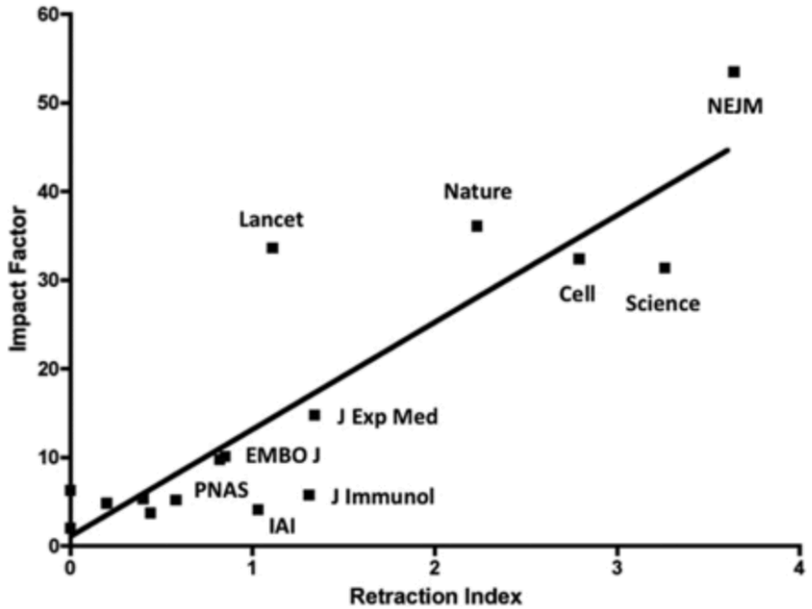
- Psychological Science (PSCI)
- Journal of Personality and Social Psychology (JPSP)
- Journal of Experimental Psychology: Learning, Memory, and Cognition

Journal	*N*	% sig. *p*	% subj. rep.
JPSP	31	23	25
JEP:LMC	27	48	54
PSCI—Cognitive	24	53	53
PSCI—Social	15	29	32
Overall	97	36	39

If this is the case, we need a much more robust culture of replication and conceptual replication, and much more tentative conclusions

INCENTIVE STRUCTURES

Careers and promotions benefit from "high profile" work



In medicine, a higher impact factor is associated with higher likelihood of retraction

Fang, Casadevall, & Morrison, 2011

often, what makes something high profile is that it's surprising — but this may be precisely the kind of thing that is less likely to replicate and less likely to be true

INCENTIVE STRUCTURES

Careers and promotions are often driven by statistics like the *h*-Index (you have published *h* papers that have been cited *h* times)

Citations increased by:

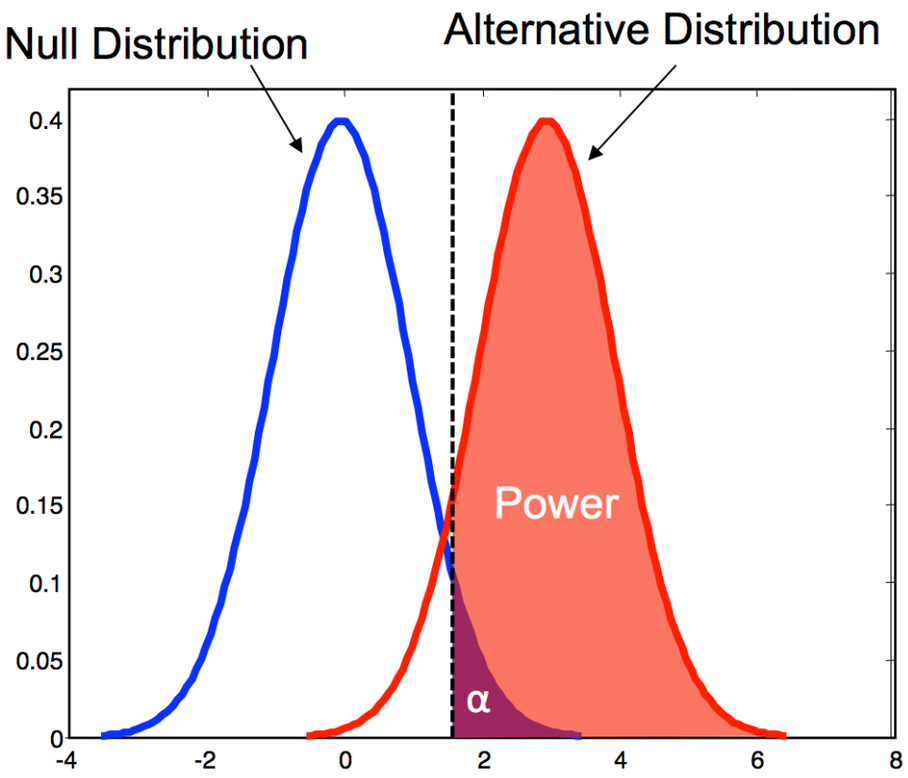
- Publishing and self-citing a lot
- Making "surprising" claims
- Publishing in high-impact journals

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Begley & Ellis, 2012

INCENTIVE STRUCTURES

Fundamentally, a problem of power: Power is the probability of (correctly) rejecting H_0 when the alternative is true



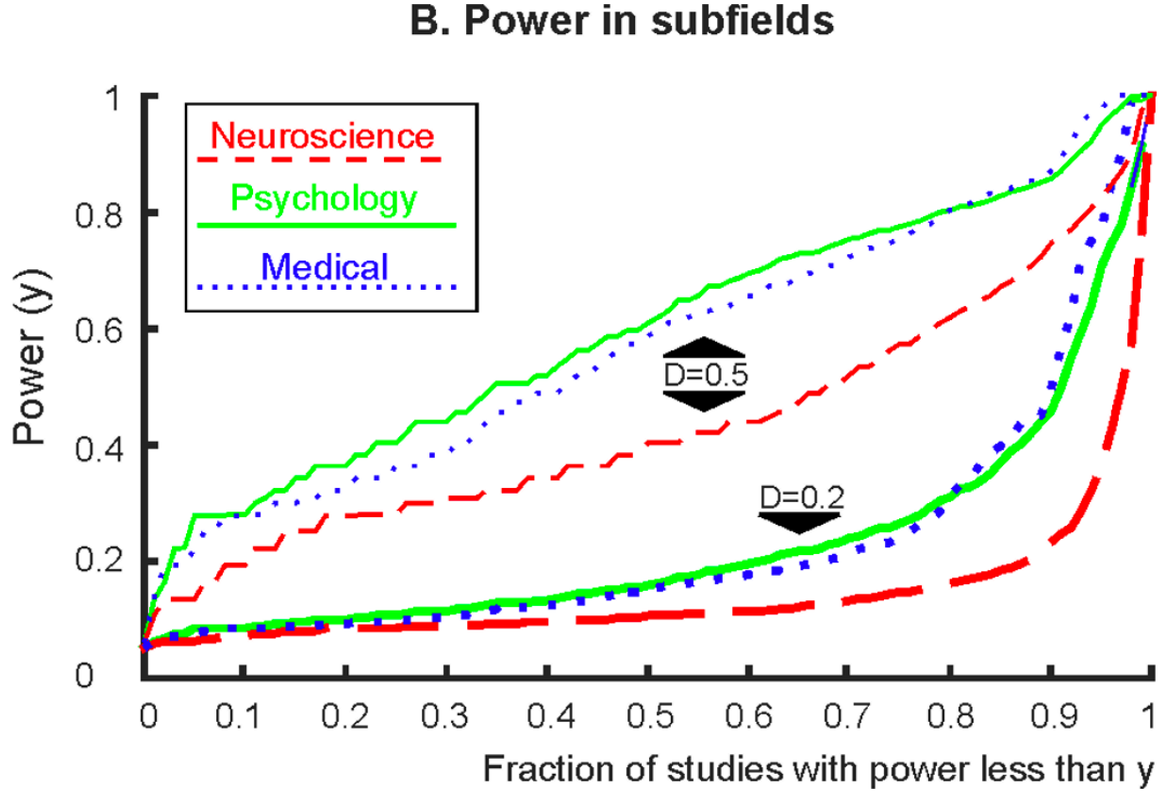
Depends on:

- Effect size (which depends on standard deviation and effect magnitude)
- Sample size
- Choice of α

Most people recommend we achieve a power of 80% (which means that if there are five "real" effects, we would find four of them)

INCENTIVE STRUCTURES

It's easier and cheaper to run underpowered studies (small sample size). In some areas it's very hard to *not* run underpowered research.



Szucs & Ioannidis, 2017

INCENTIVE STRUCTURES

When studies are underpowered, it is also likely that reported statistically significant effects are larger than they really are: this is called the **winner's curse**

INCENTIVE STRUCTURES

This is one aspect of publication bias and the file-drawer effect

Findings are more likely to get rejected if they aren't significant

(not entirely irrational: there are many (often uninteresting) reasons that something might not "work")

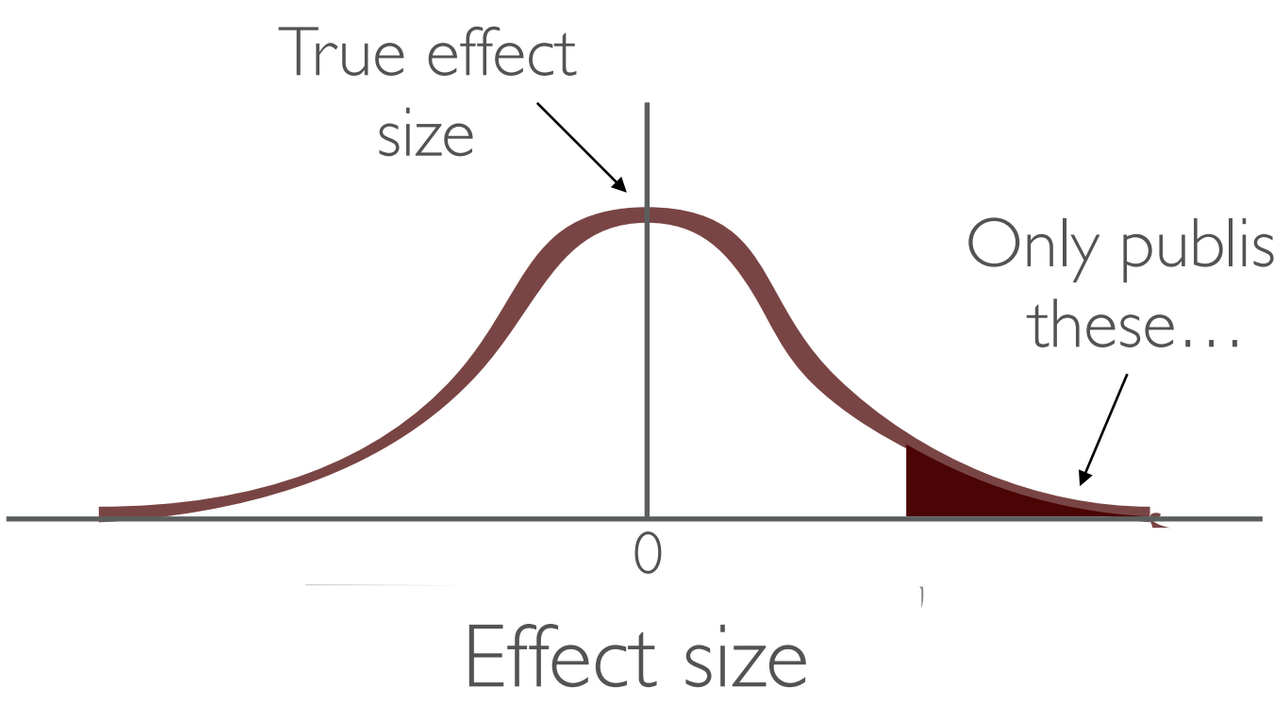
People know this and often don't even try

Leads to a biased sampling of outcomes!

STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Null hypothesis significance testing with thresholds



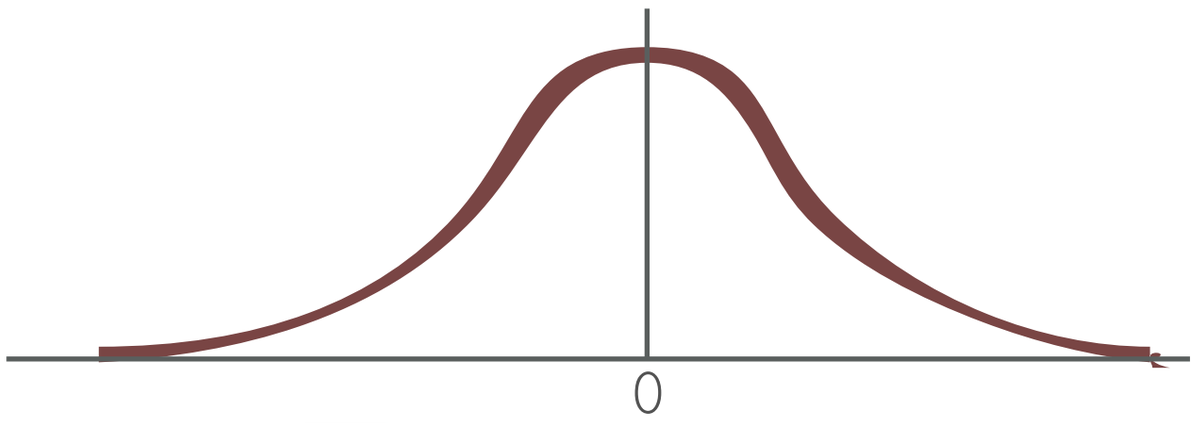
… so it looks like there is an effect in the literature

STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Null hypothesis significance testing with thresholds

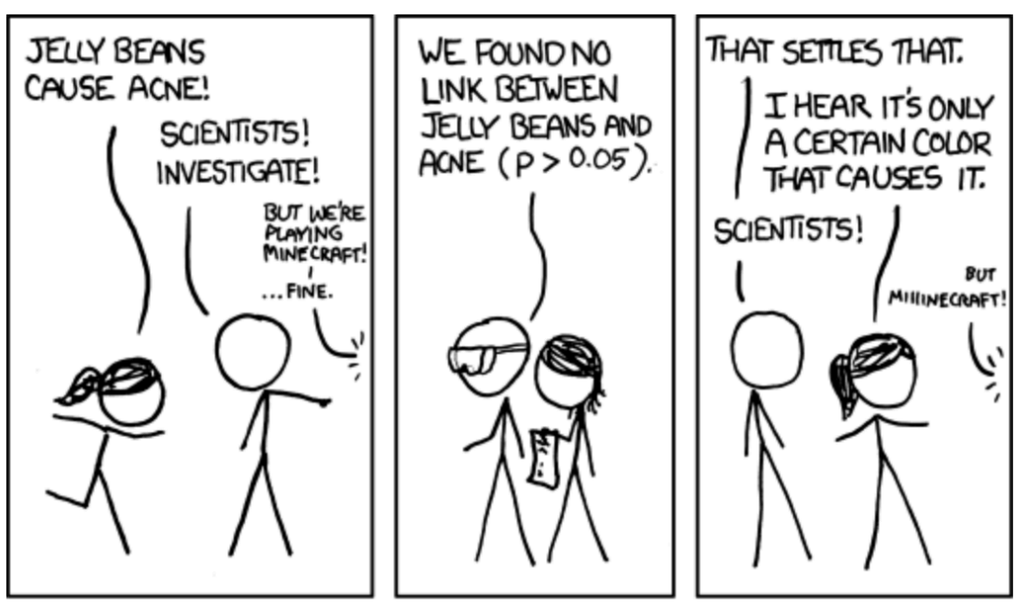
Need to not threshold as a criterion for publication

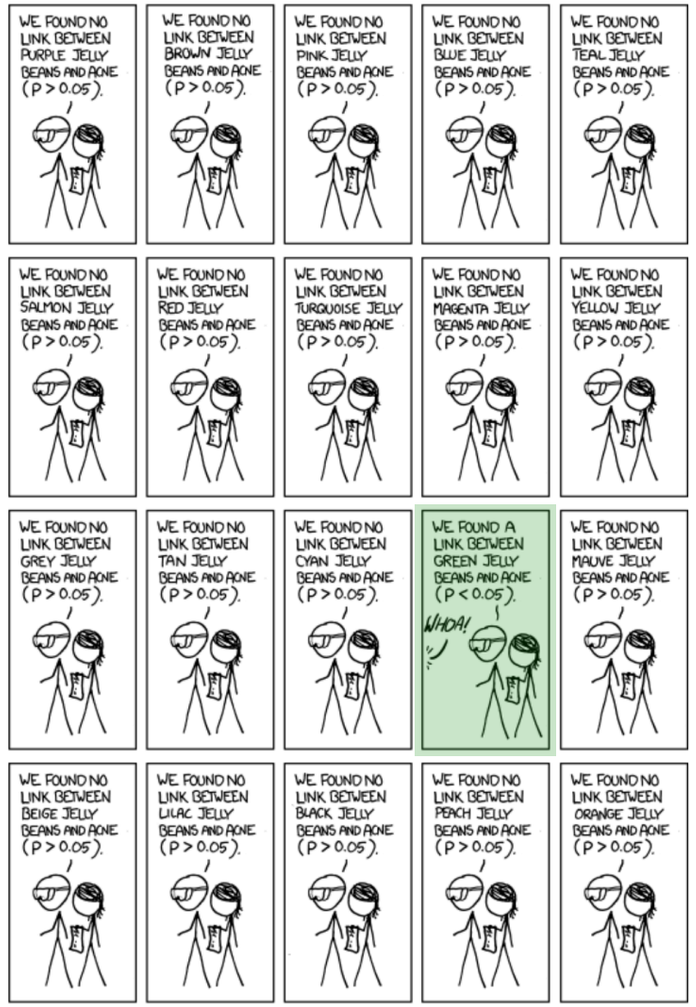


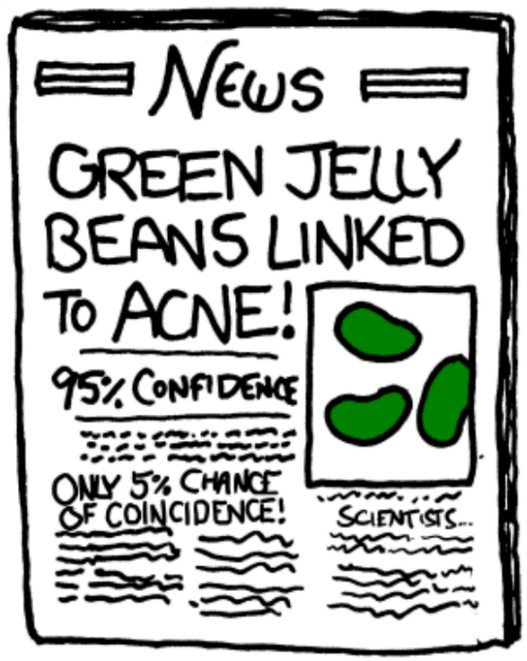
STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Too many experimenter degrees of freedom, too many tests that could be (and are) run



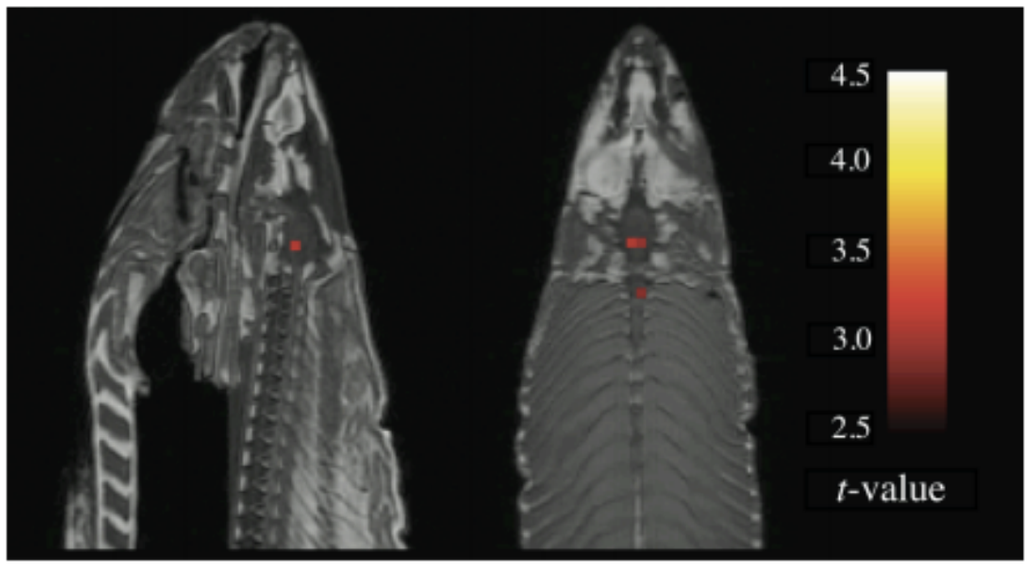




STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Too many experimenter degrees of freedom, too many tests that could be (and are) run



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Too many experimenter degrees of freedom, too many tests that could be (and are) run
  - Outlier removal
  - Subsets of trials or participants
  - Which variables to look at
  - Which tests to run
  - Defaults / assumptions of tests
  - What the hypotheses are

STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

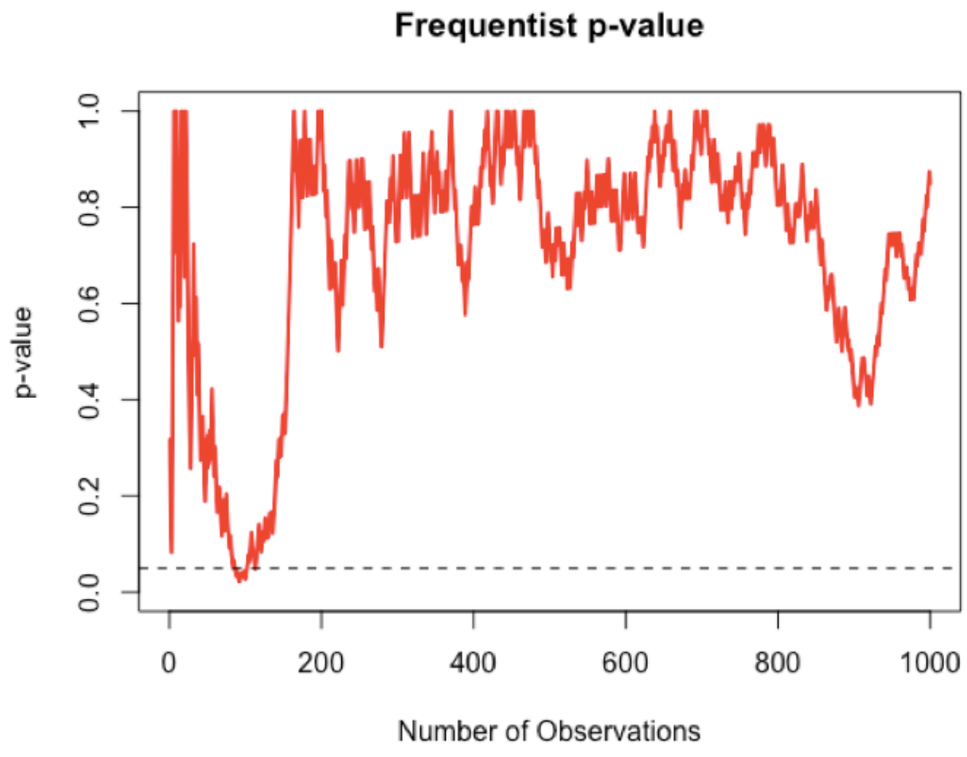
- Keep running new participants until you have an effect, then stop

open R program

`demonstration.R`

Flips a coin n times. After each flip, uses a chi-squared test to decide if the data so far is significantly different from chance.

STOP testing when it is



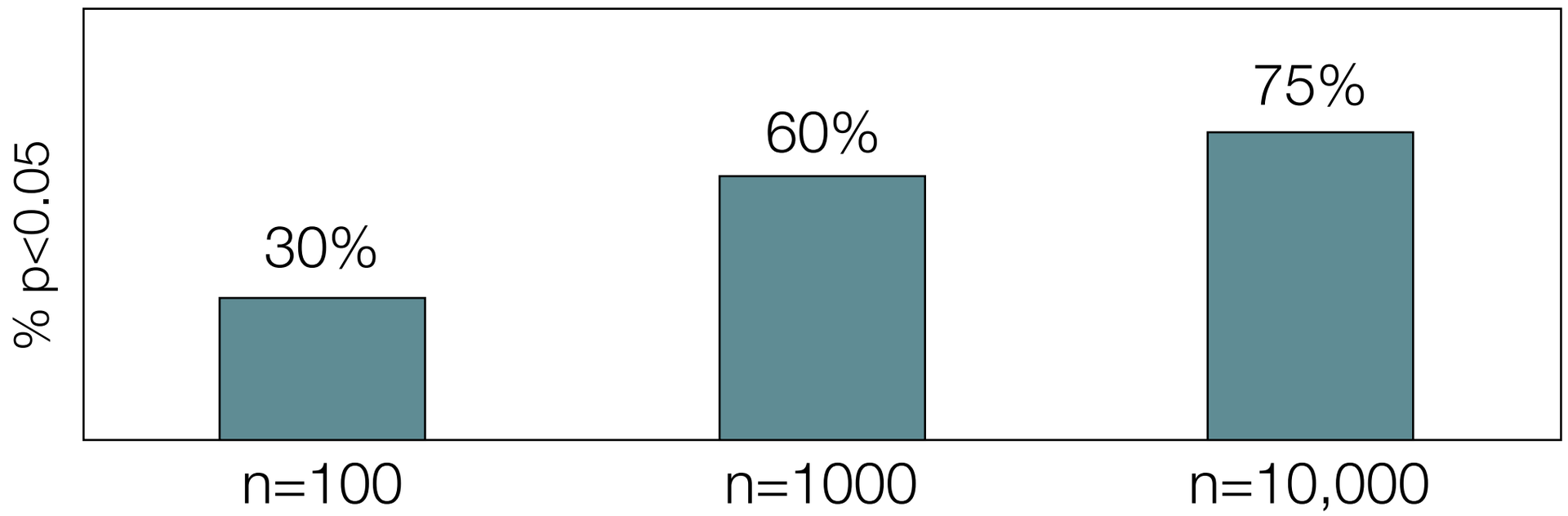
`sequentialTestingFrequentist(n=1000)`

STATISTICAL PRACTICE

Many of the problems derive from how we do our research…

- Keep running new participants until you have an effect, then stop

How many times do you incorrectly find $p<0.05$ at each sample size?



STATISTICAL PRACTICE

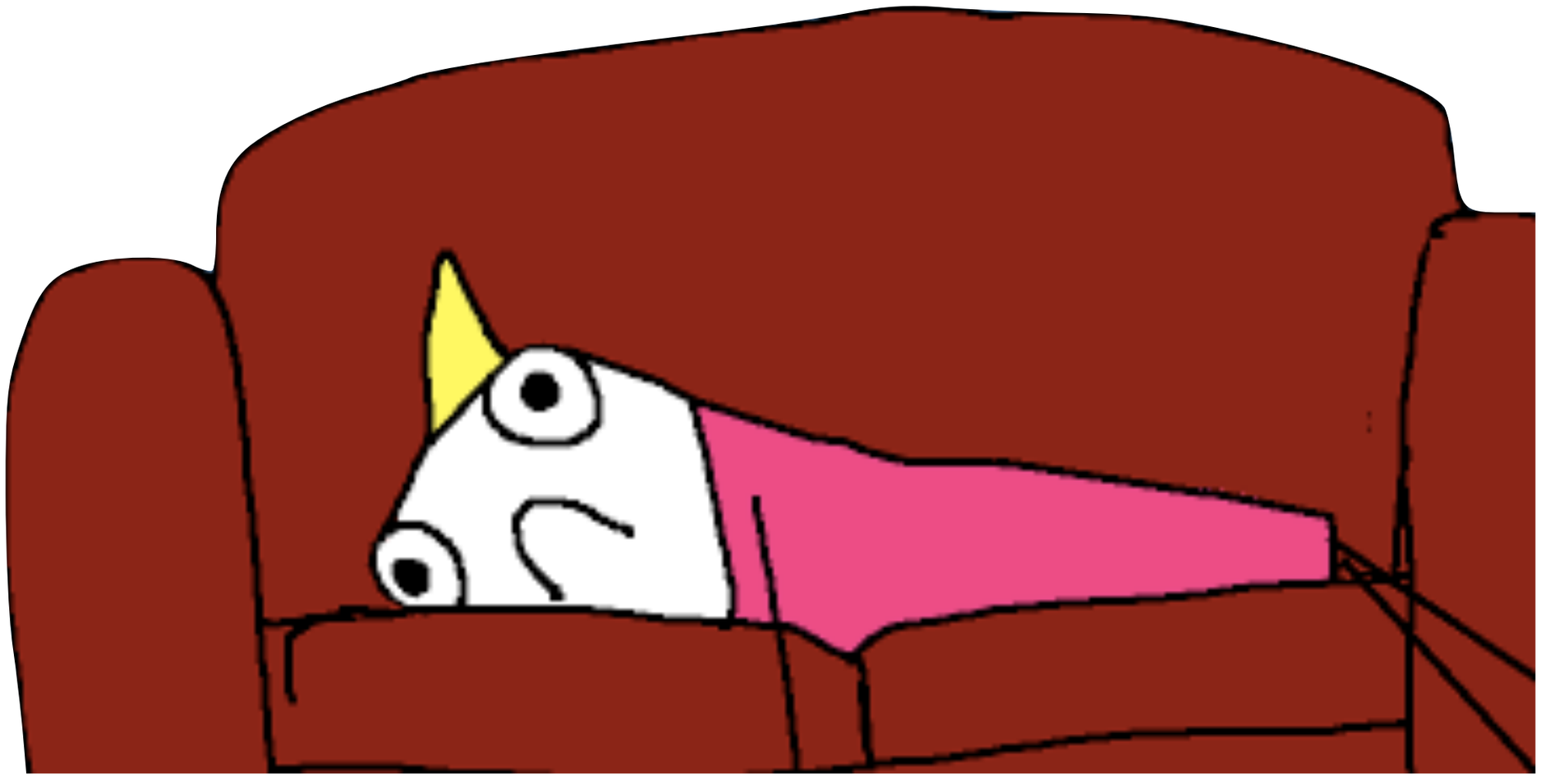
Many of the problems derive from how we do our research…

- Not replicating

Many of these problems could be avoided if we routinely replicated findings, but there is no robust culture of doing so: few career incentives to, plus it's more boring.

Only 3% of 1151 psychology journals explicitly state in their aims or instructions that they accept replications (Martin & Clarke, 2017)

DOES THIS MEAN ALL IS LOST?



NOT NECESSARILY…

I actually think this has been pretty good for psychology (and in the long run will be even better).

- Much larger focus on good methods
- Improved practices, better data, more robust findings

SOLUTIONS

- Increased sample size

Investigating variation in replicability: A “many labs” replication project.

EXPORT ★ Add To My List Database: PsycARTICLES Journal Article

Klein, Richard A. Ratliff, Kate A. Vianello, Michelangelo Adams Jr., Reginald B. Bahník, Štěpán Bernstein, Michael J. Bocian, Konrad Brandt, Mark J. Brooks, Beach Brumbaugh, Claudia Chloe Cemalcilar, Zeynep Chandler, Jesse Cheong, Winnee Davis, William E. Devos, Thierry Eisner, Matthew Frankowska, Natalia Furrow, David Galliani, Elisa Maria Hasselman, Fred Hicks, Joshua A. Hovermale, James F. Hunt, S. Jane Huntsinger, Jeffrey R. IJzerman, Hans John, Melissa-Sue Joy-Gaba, Jennifer A. Barry Kappes, Heather Krueger, Lacy E. Kurtz, Jaime Levitan, Carmel A. Mallett, Robyn K. Morris, Wendy L. Nelson, Anthony J. Nier, Jason A. Packard, Grant Pilati, Ronaldo Rutchick, Abraham M. Schmidt, Kathleen Skorinko, Jeanine L. Smith, Robert Steiner, Troy G. Storbeck, Justin Van Swol, Lyn M. Thompson, Donna van ‘t Veer, A. E. Vaughn, Leigh Ann Vranka, Marek Wichman, Aaron L. Woodzicka, Julie A. Nosek, Brian A.

Citation

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), 142-152. http://dx.doi.org/10.1027/1864-9335/a000178

Abstract

Although replication is a central tenet of science, direct replica
replicability of 13 classic and contemporary effects across 36
effects replicated consistently. One effect—imagined contact
effects—flag priming influencing conservatism and currency p
compared whether the conditions such as lab versus online o
large they did not. The results of this small sample of effects s
the sample and setting used to investigate the effect. (PsycIN

manybabies.github.io

ManyBabies website

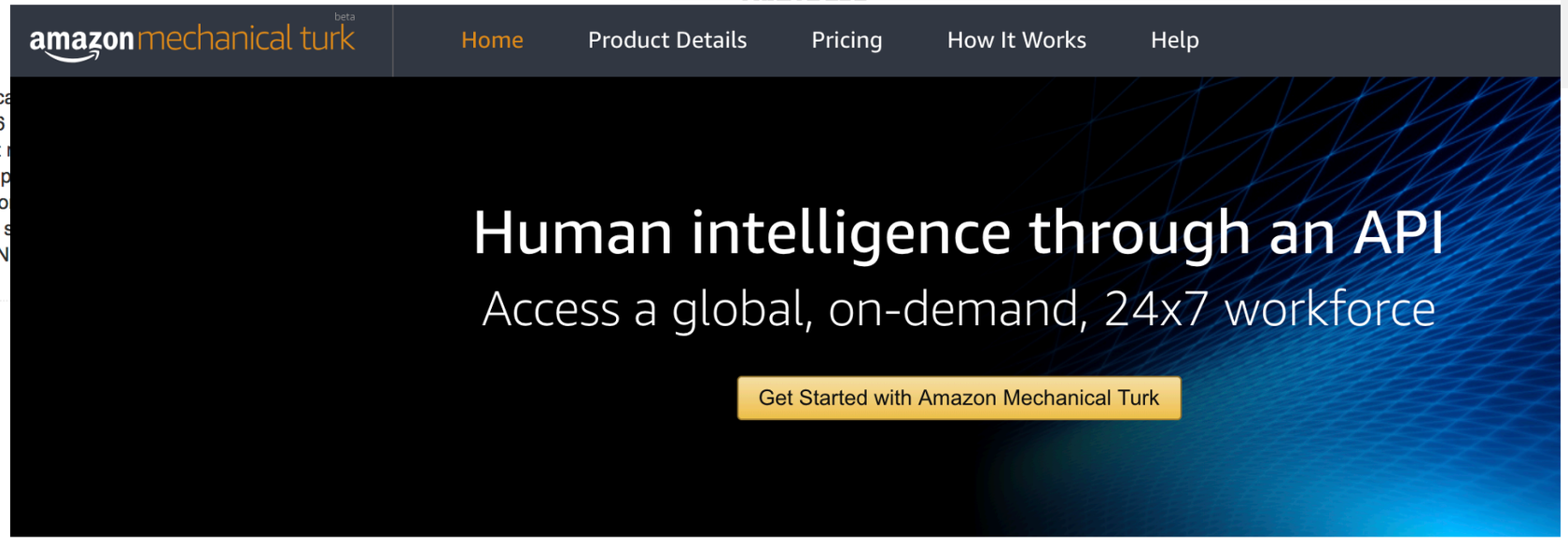
The ManyBabies Project

ManyBabies is a collaborative project for replication and best practices in developmental psychology research. Our goal is to bring labs together to address difficult outstanding theoretical and methodological questions about the nature of early development and how it is studied.

News

- **Starting January 2018, we are accepting applications for a postdoctoral fellowship!**
- We hosted a workshop at the Cognitive Development Society in Fall 2017.
- We have a special issue in Infant Behavior and Development, due date April 1, 2018, on the topic of “Replication, Collaboration, and Best Practices in Infancy Research”

Projects



SOLUTIONS

- Increased sample size

- Support for replications (and discussion of how to further improve)

Behav Brain Sci. 2017 Oct 25:1-50. doi: 10.1017/S0140525X17001972. [Epub ahead of print]

MAKING REPLICATION MAINSTREAM.

Zwaan RA[1], Etz A[1], Lucas RE[1], Donnellan MB[1].

Author information

Abstract

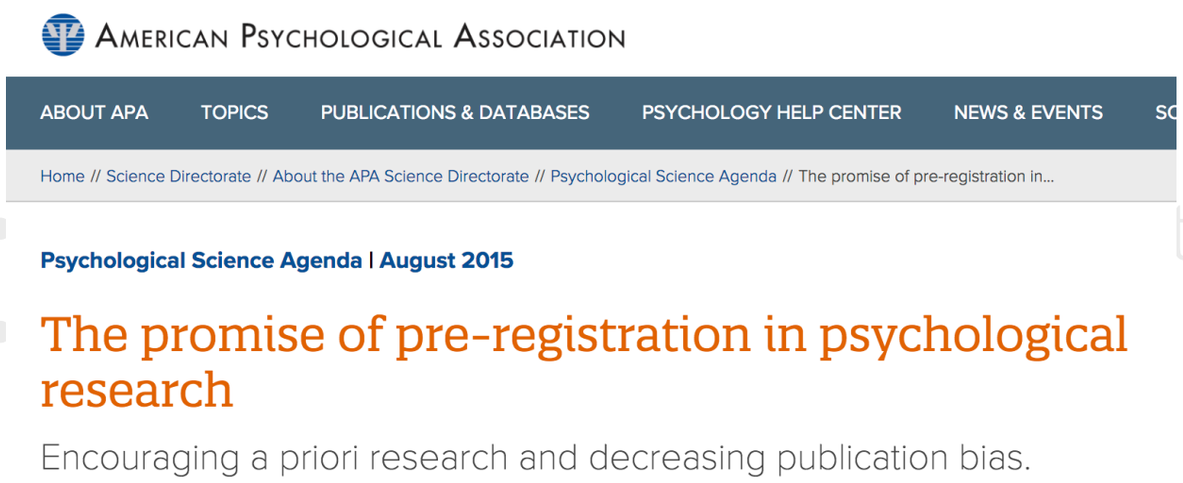
Many philosophers of science and methodologists have argued that the ability to repeat studies and obtain similar results is an essential component of science. A finding is elevated from single observation to scientific evidence when the procedures that were used to obtain it can be reproduced and the finding itself can be replicated. Recent replication attempts show that some high profile results---most notably in psychology, but in many other disciplines as well---cannot be replicated consistently. These replication attempts have generated a considerable amount of controversy and the issue of whether direct replications have value has, in particular, proven to be contentious. However, much of this discussion has occurred in published commentaries and social media outlets, resulting in a fragmented discourse. To address the need for an integrative summary, we review various types of replication studies and then discuss the most commonly voiced concerns about direct replication. We provide detailed responses to these concerns and consider different statistical ways to evaluate replications. We conclude there are no theoretical or statistical obstacles to making direct replication a routine aspect of psychological science.

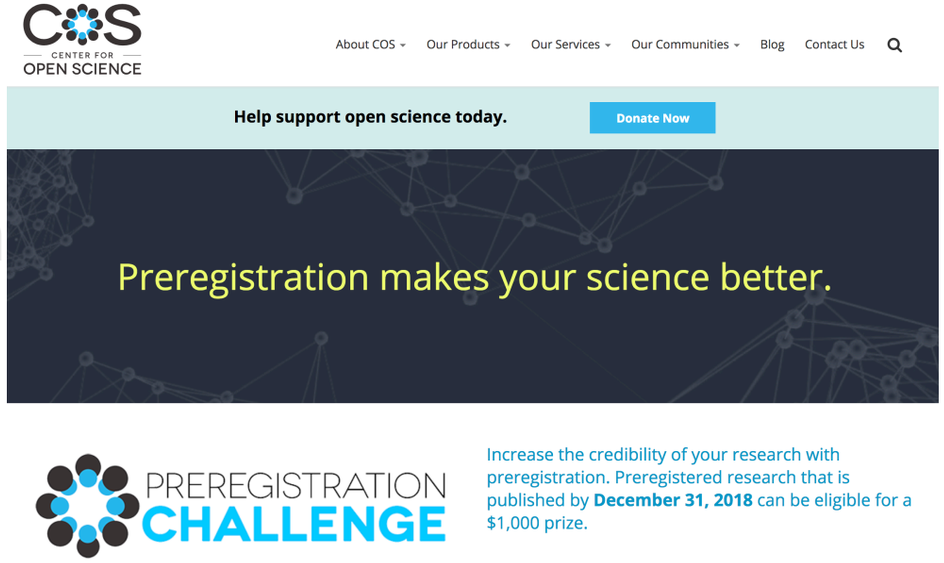
SOLUTIONS

- Increased sample size

- Support for replications (and discussion of how to further improve)

- Pre-registration of studies and public datasets





SOLUTIONS

- Increased sample size
- Support for replications (and discussion of how to further improve)
- Pre-registration of studies and public datasets
- More Bayesian statistics, testing the alternative, modelling

ULTIMATELY

statistics is helpful, and necessary.
but it's not a panacea.

ultimately any tool is only as good as
the people who wield it

THE GOAL OF THIS SCHOOL

To give you some of the tools you need to do good science