

MAS.573 Moving Beyond the Replication Crisis

How to Spot Misleading Social Science and Design Better Experiments

I am thrilled to see so many people here. In my opinion, there is everything at stake when it comes to social psychological. If we want to answer questions or build technologies that make their lives meaningful, that help build or break habits, that help us understand mental health, that preserve our democracy and public discourse. All of these questions rely on us understanding individual psychology and the influence of the environment on it.

We've really been getting it wrong for the last 20-30 years, and I think that cross-disciplinary places like right here, where we can combine rigorous statistics with sensors that measure things about people's lives and design interesting interventions, that the future of empirical psychology will happen.

MAS.S73: Moving Beyond the Replication Crisis

Lecture 1: Introduction and Overview

David Ramsay



Imagine that it's 2011. You're the editor of a well respected social psychology journal the Journal of Personality and Social Psychology. and you have a big problem.

And you receive a paper from this professor— Daryl Bem, from Cornell University— he has submitted a paper summarizing two years of work, his career's magnum opus. It follows all of the right methods. it's rigorously done. it looks just like all the other papers you accepted. and you really, really don't want to publish it.



and that's because this paper is proved that all of us are capable of seeing the future, retrocausation or psi. it includes many different experiments that prove that this is a real phenomena. Daryl Bem Really believes in this phenomena.

Bem, Daryl J. "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." Journal of personality and social psychology 100.3 (2011): 407. (STILL NOT RETRACTED)

published with editorial: <https://psycnet.apa.org/record/2011-01911-001>

<https://replicationindex.com/2018/01/05/bem-retraction/>

Bem, Daryl, et al. "Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events." F1000Research 4 (2015).

<https://replicationindex.com/page/24/?wref=bif>



53% of the time people chose pornographic image, and 57% of the time for 'stimulus seekers'. More than random!

What would you do with this paper?

"After a rigorous review process, involving a large set of extremely thorough reviews by distinguished experts in social cognition, we are publishing the following article by Daryl J. Bem... To some of our readers it may be both surprising and disconcerting that we have decided to publish Bem's article.... **We openly admit that the reported findings conflict with our own beliefs about causality and that we find them extremely puzzling.** Yet, as editors we were guided by the conviction that this paper—as strange as the findings may be—should be evaluated just as any other manuscript on the basis of rigorous peer review..."

-Editorial published by Journal Editors

The New York Times

Journal's Paper on ESP Expected to Prompt Outrage

Facebook Twitter YouTube Plus Comment



By Benedict Carey
Jan. 5, 2015

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the paper, to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

well the journal editors published it, and alongside of it they wrote a small editorial which you can see here on the left.

The next day, it was a new york times front page story.

'one of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events. The decision is already mortifying scientists, generating a mixture of amusement and scorn.'

Boon for the field. Shone a light on shoddy methods. Thank goodness for this principled decision.



An Open Letter to John Bargh
2012

As all of you know, of course, questions have been raised about the robustness of priming results.... your field is now the poster child for doubts about the integrity of psychological research... people have now attached a question mark to the field, and it is your responsibility to remove it... all I have personally at stake is that I recently wrote a book that emphasizes priming research as a new approach to the study of associative memory...Count me as a general believer... My reason for writing this letter is that **I see a train wreck looming.**

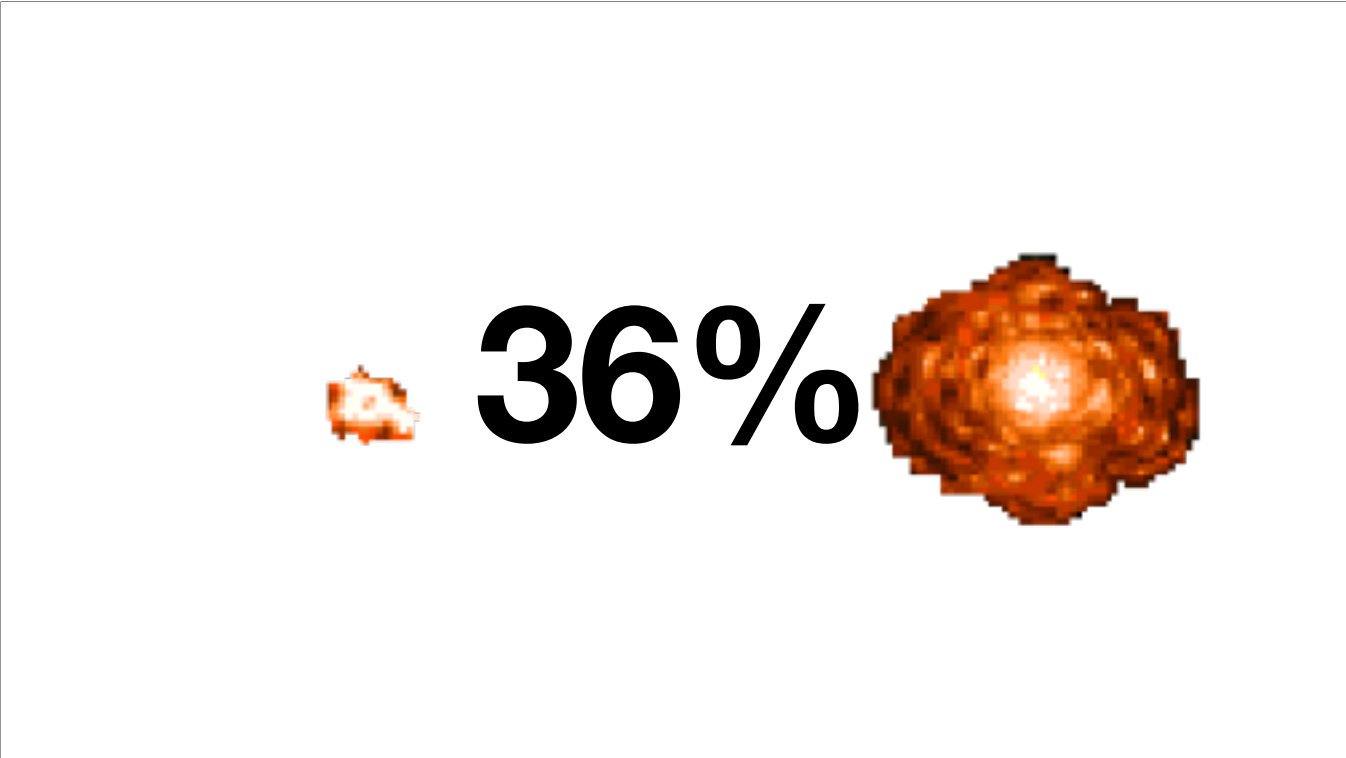
<https://www.nature.com/articles/nature.2012.11535>

in the wake of failed replication 2012 by Doyen of social priming— the idea that subtle influences in the environment drive behavior meaningfully— Daniel Kahneman- nobel prize winner, top psychologist, and author of thinking fast and slow— wrote an open letter suggesting that there was a train wreck looming for the field.

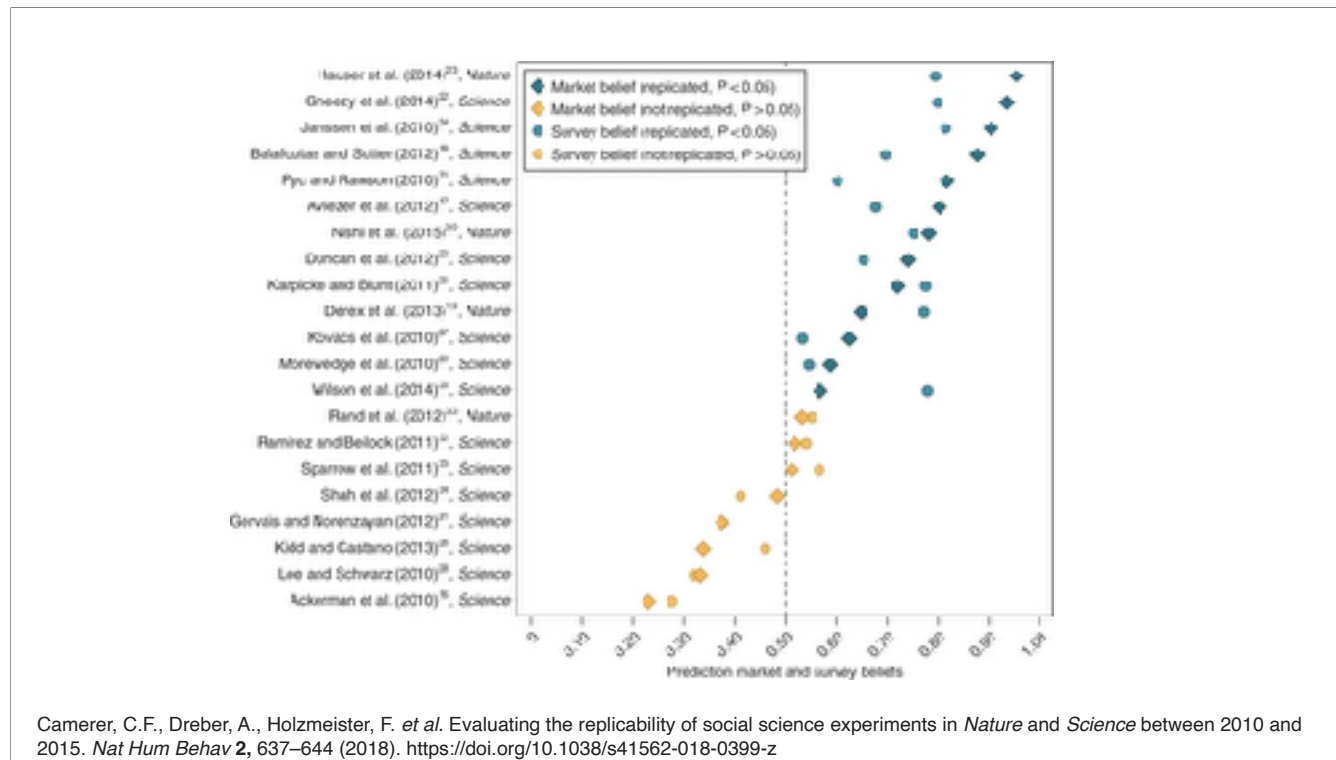
Journal	% Replicated
Journal of Experimental Psychology: Learning, Memory, and Cognition	48
Journal of Personality and Social Psychology (<i>Social Psych Papers Only</i>)	23
Psychological Sciences (<i>Social Psych Papers Only</i>)	29
Psychological Sciences (<i>Cognitive Psych Papers Only</i>)	53
Overall	36

A reproduction of Table 1. from the Open Science Collaboration's "[Estimating the reproducibility of psychological science.](#)" *Science* 349.6251 (2015). Notice that Social Psychology papers specifically fare far worse than others. Based on 100 replication attempts carried out by 270 different authors. This study was [criticized by Harvard's Dan Gilbert](#) and responded to by OSC's [Nosek](#) among [others](#).

And he was right. Wow. That's not 50/50, there is information here. It's *actually less likely* to be true if it's published; if you know nothing else, you should actually become more doubtful about a reported hypothesis.



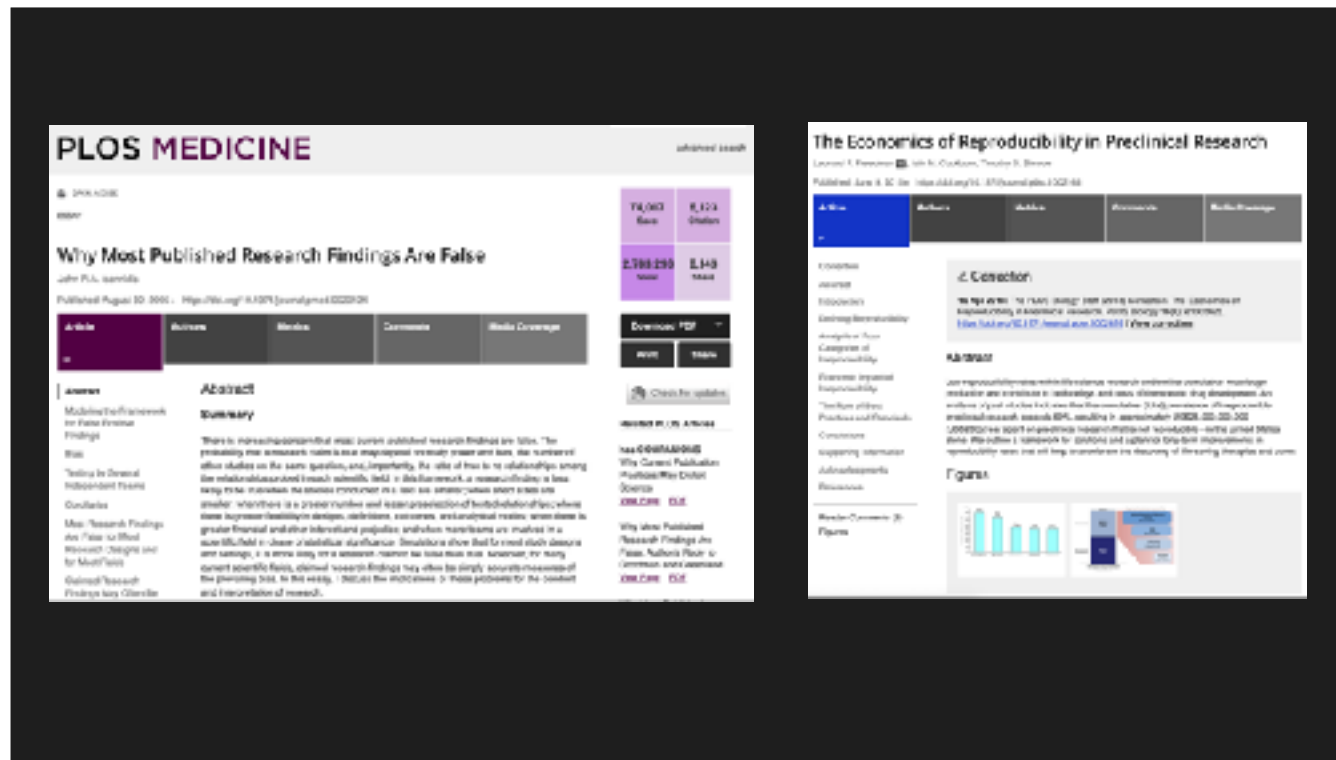
let's pause and appropriately let this sink in.



21 higher power replications, with sample sizes on average about five times higher than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. the estimated true-positive rate is 67% in a Bayesian analysis. Things are better for Science and Nature.

yellow didn't replicate. blue did.

Human predictions are good! Either we have really good intuition about human behavior, or we're really good at reading studies and figuring out when their methodology is poor, or both.



Not just social psychology, these issues are core to many fields. Medical research, particularly pharmacological and pre-clinical cancer research.

2005 Ioannidis (Stanford) article. Famous, kicked off a debate, heavily questioned.

<https://pubmed.ncbi.nlm.nih.gov/24068246/>

<https://replicationindex.com/2020/12/24/ioannidis-is-wrong/>

top two retraction frauds are Anesthesiologists Yoshitaka Fujii (183 papers) and Joachim Boldt (153).

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John F.A. Janszids, MD

CLINICAL RESEARCH ON important questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be biased by large ones.^{1,2}

Recently, there is more evidence on disagreements between epidemiological studies and randomized trials.³⁻⁷ Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives most attention and dominates science.

Background Controversy and uncertainty arose when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discover whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 2 major general clinical journals or high-impact-factor specialty journals in 1990-2008 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (14%) had found effects that were stronger than those of subsequent studies, 23 (47%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly cited nonrandomized studies had more optimistic and/or had found stronger effects ($P < .001$) of 39 randomized control trials ($P = .008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P = .049$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with “negative” results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provide contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

www.ama-assn.org

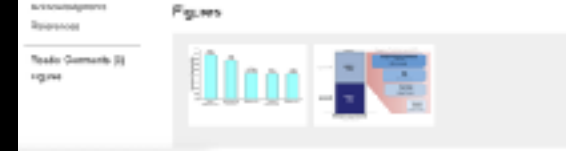
The Economics of Reproducibility in Preclinical Research

Author: Michael S. H. H. Koozekan, M.D., Ph.D. | Published: June 8, 2011 | DOI: 10.1093/bioinformatics/btt103

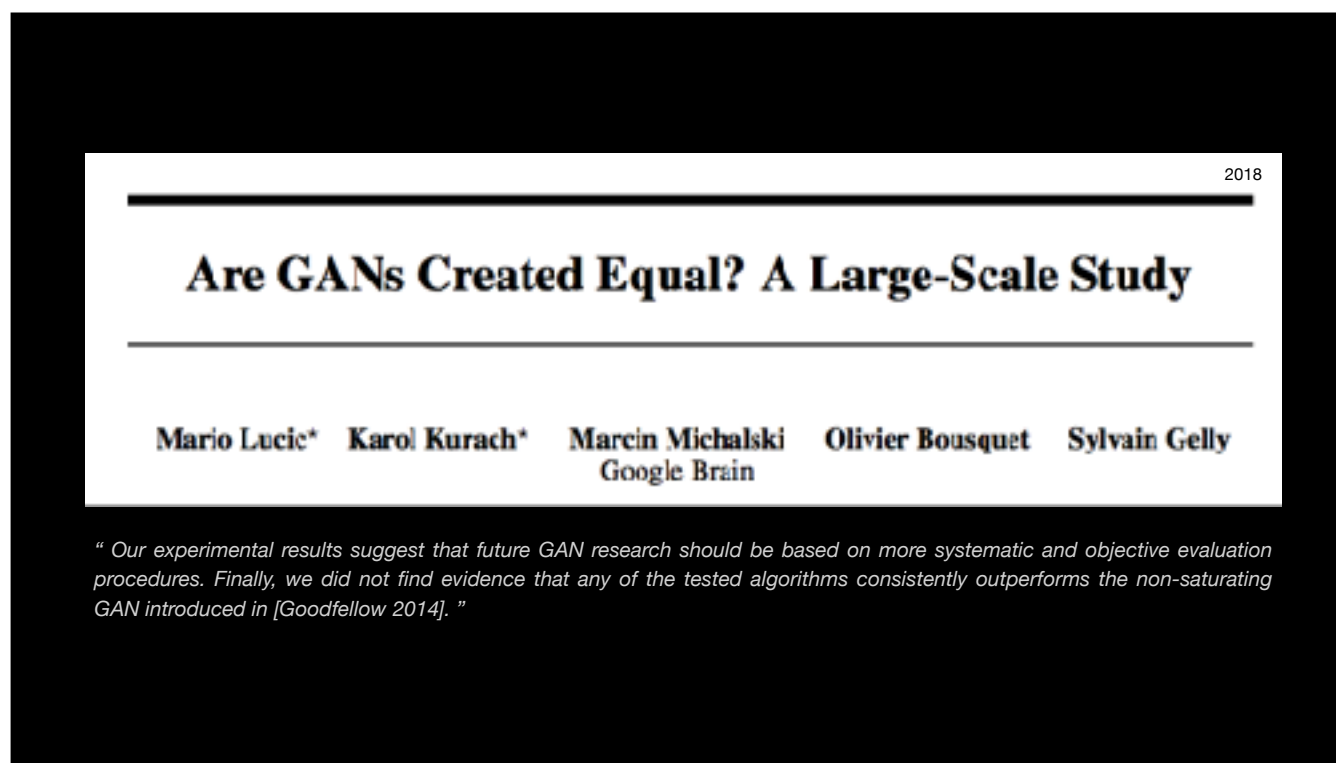
Article	Authors	Article	Comments	Media Coverage
---------	---------	---------	----------	----------------

4. Correction
John S.M. T. van Fliet, Ed. | DOI: 10.1093/bioinformatics/btt103

Abstract
Low reproducibility means unreliable science research outcomes. Consequently, large pharmaceutical and academic labs and other stakeholders that depend on the analysis of publication outcomes face the cumulative financial consequences of irreproducible preclinical research results. Results in quantitative research are more likely to be reproduced than qualitative research results. In the United States alone, the public is expected to invest and a private industry investments in reproducibility data that will help to accelerate the discovery of new drugs and other products.



50 highly cited, major impact studies in literature— 7/34 didn't replicate, 7/34 weaker effect than reported.



Oct 2018. 2014 Ian Goodfellow GAN neural network is still the best over multiple datasets and hyper-parameters, despite many years and claims of improvement. Overfitting and sampling issues at the system level.

“Our experimental results suggest that future GAN research should be based on more systematic and objective evaluation procedures. Finally, we did not find evidence that any of the tested algorithms consistently outperforms the non-saturating GAN introduced in [9].”

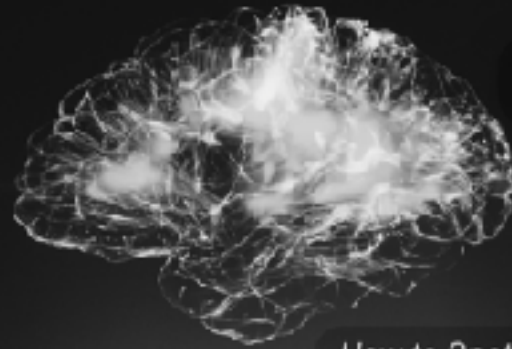


Also FMRI, issues include low-power tests, many different analysis techniques.

2020 fMRI data (NARPS study) The analysis of a single fMRI dataset by 70 independent analysis teams, all of whom used different analysis pipelines, revealed substantial variability in reported binary results, with high levels of disagreement across teams for most of the tested hypotheses. For every hypothesis, at least four different analysis pipelines could be found that were used in practice by research groups in the field and resulted in a significant outcome.

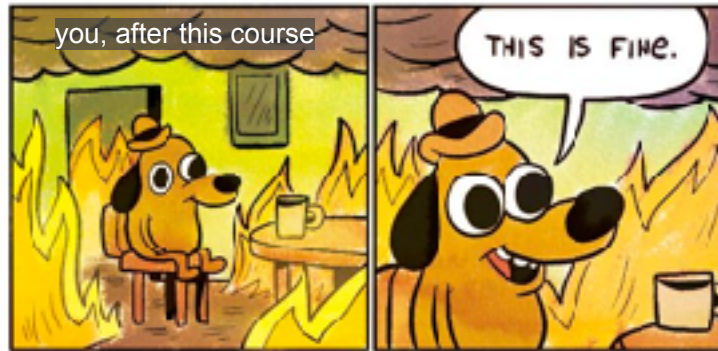
prediction market on fMRI, not nearly as good as behavior, general overestimate of significance.

fMRI average power is estimated between 8 and 31%.



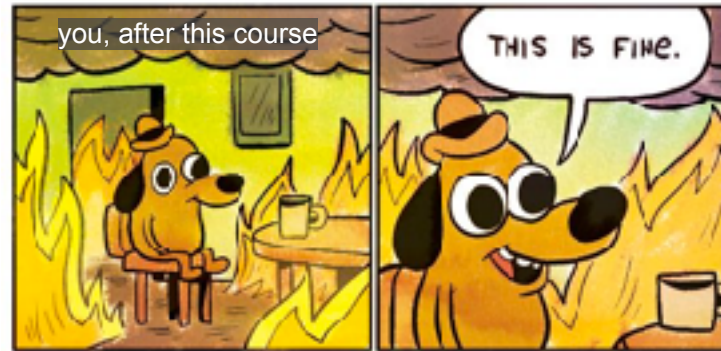
MAS.573 Moving Beyond the Replication Crisis

How to Spot Misleading Social Science and Design Better Experiments



Why this Class?

For me this topic is really a personal one. I'm in the resenv group here, and I really bought into a lot of the social psychology research. A lot of the way I viewed the world was really wrong— this idea that subtle changes in the environment control and puppeteer behavior in really profound ways. That's a major misconception we will talk about— I lost time to this, I see others lose time to it.



Why this Class?

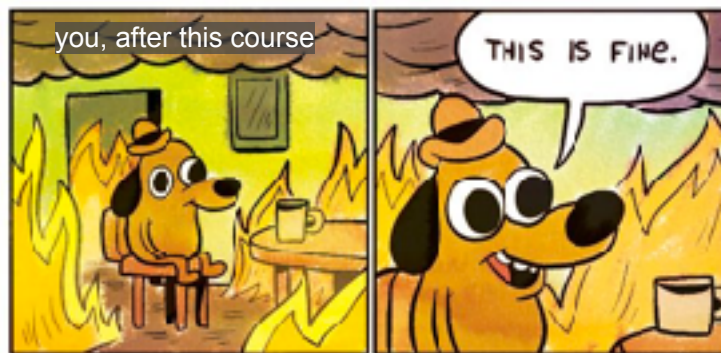
- Hard to find centralized resources
 - decentralized, blogs (!)
 - professionally charged
 - affects lots of disciplines
 - evolving literature and field of meta-statistics
- Few cover the full extent of the crisis
- Few focus on solutions for individual researchers

what this class is and why:

- replication crisis, vague and underspecified
- no central resource, controversial
- critique common; what do we know/what is accurate view?
- how to move forward? Bayesian reasoning, computational cognitive science, philosophy of science/logic/epistemology
- HCI researchers and computational social scientists are in a unique position

what we will cover and expectations:

- the real victims here are the scientists we will be talking about. Their careers, reputations, and livelihoods are in jeopardy because they did what they were taught and they didn't have a sophisticated understanding of statistics. and it's easy to see the hypocrisy now, and be slightly bitter about it. Don't give in to the natural Schadenfreude
- it's really easy to inadvertently p-hack, be understanding



Why this Class?

- Assignment and Structure
- Tone in class
- Difficulty of Statistics

1/11 Intro to Crisis (David)

1/13 Practical Skills for Navigating the Crisis (Noah)

1/20 Philosophical Groundwork: Induction and Causal Reasoning (Matt)

1/25 New Tools and Moving Forward: Contextualized Causal Reasoning (David)

1/27 Practical Stories of the Replication Crisis (David and Students)

I. METHODS

MISAPPLIED STATISTICS

P-Hacking, Publication Bias, File Drawer Effects, HARKing, QRPs, Researcher DOF

II. SYSTEMS

SYSTEMIC INCENTIVES

Fraud, Hype, Motivated Misinterpretation, Overgeneralization, Narrative Support, Status Quo Bias

III. ONTOLOGIES

FUNDAMENTAL ASSUMPTIONS

Psychometrics, Taxometrics, Analytic/Gestalt, Idiographic/Nomothetic, Statistical/Causal Reasoning

how to conceive of the crisis; we'll talk about 1/2 today. 3 is really important and often addressed separately, but we'll talk about it in later lectures.

Conceptual Stats 101

so to have this conversation at all, we do need to talk about some basic statistical ideas and understand them at a conceptual level. So we'll start with a review of this with just a simple two sample t-test; hopefully this is review, if it's not, don't worry, I will try to make it possible to follow along even if these ideas are new or foreign.

*In every study reviewed, **the majority of researchers (56%–97%)** exhibited one or more of these [statistical] delusions. Psychology departments need to begin teaching **statistical thinking, not rituals.***

General Article

**Statistical Rituals: The Replication
Delusion and How We Got There**

Geert Gigerenzer

Hertie Center for Risk Literacy, Max Planck Institute for Human Development, Berlin, Germany

aps
ADVANCED
PSYCHOLOGICAL SCIENCE

Advances in Methods and
Practices in Psychological Science
2011, Vol. 2(1) 196–218
© The Author(s) 2011
http://dx.doi.org/10.1037/a0021711
aps.sagepub.com/journalsPermissions.nav
DOI: 10.1177/2154279811417150
www.psyc.sagepub.com/journals.nav

SAGE

motivation.

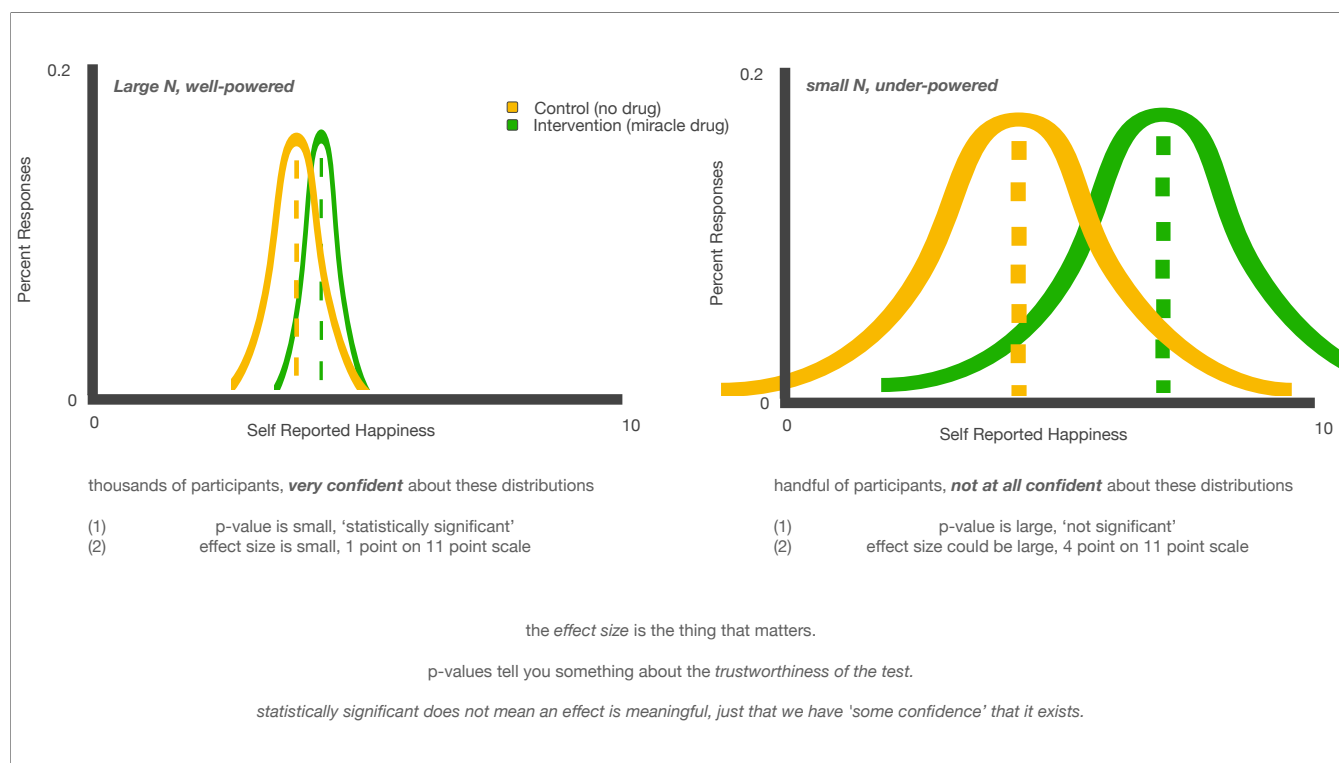
you don't have to *master* statistics in all its intricacies. You do have to understand it conceptually.

Driver of a car? No. Mechanic for the car? No. But we need an *understanding of the pieces of an engine and have a sense for how it works, what its limitations are, and when it's failing.* Nerdy Racecar Driver.



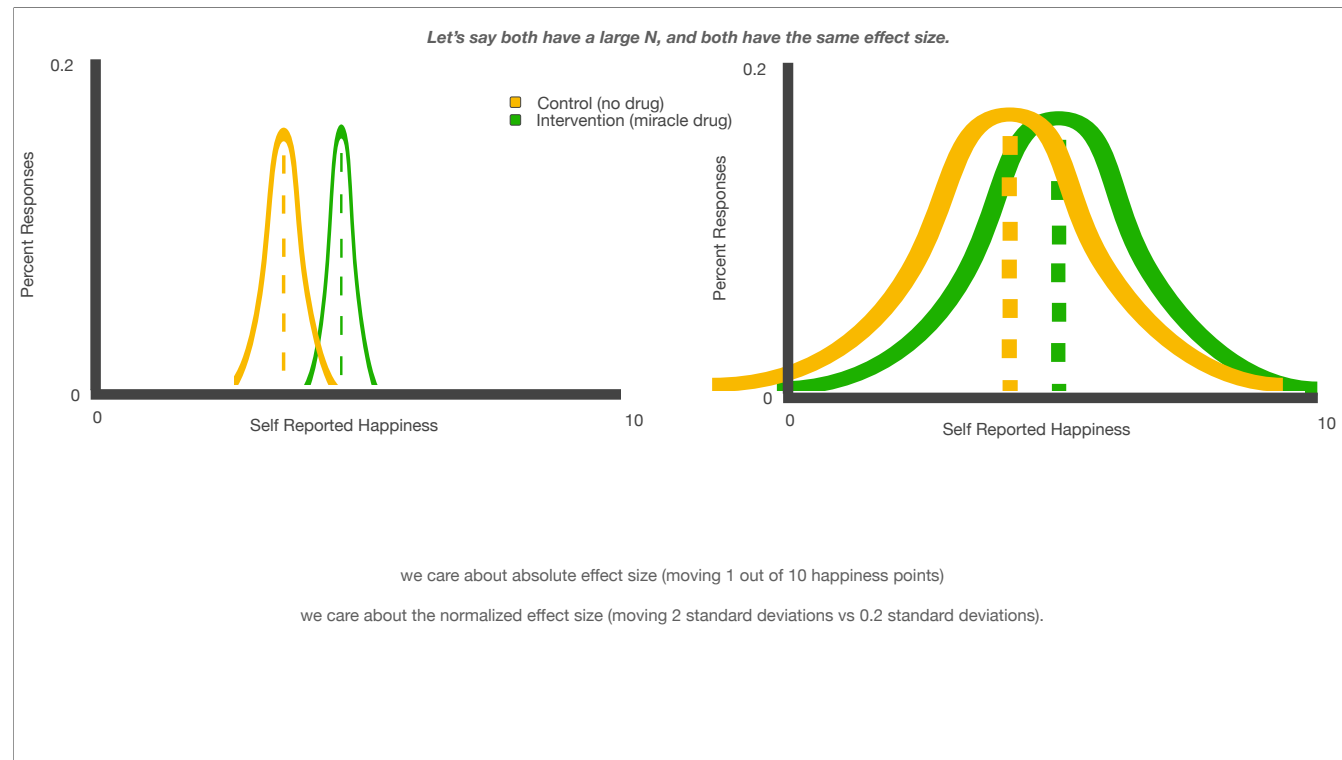
small n vs large n, small effect size vs large effect size.

effect size is generally what we care about, p-value has nothing to do with it. p-value tells us how likely are results are due to noise, randomly drawing from the same underlying distribution. Need both. Be sure to understand the difference here; effect size is really the thing that tells us if an effect is meaningful, p tells us if we sampled enough in our test to be confident in that effect.



small n vs large n, small effect size vs large effect size.

effect size is generally what we care about, p-value has nothing to do with it. p-value tells us how likely are results are due to noise, randomly drawing from the same underlying distribution. Need both. Be sure to understand the difference here; effect size is really the thing that tells us if an effect is meaningful, p tells us if we sampled enough in our test to be confident in that effect.



in this case, we taken, on one hand a group of 3s and made them 4s. We've moved them 2 standard deviations. In the second we took a group of people spread from 0 to 9 and nudged them to be 1 to 10, 0.2 std deviations.

we care about effect size.

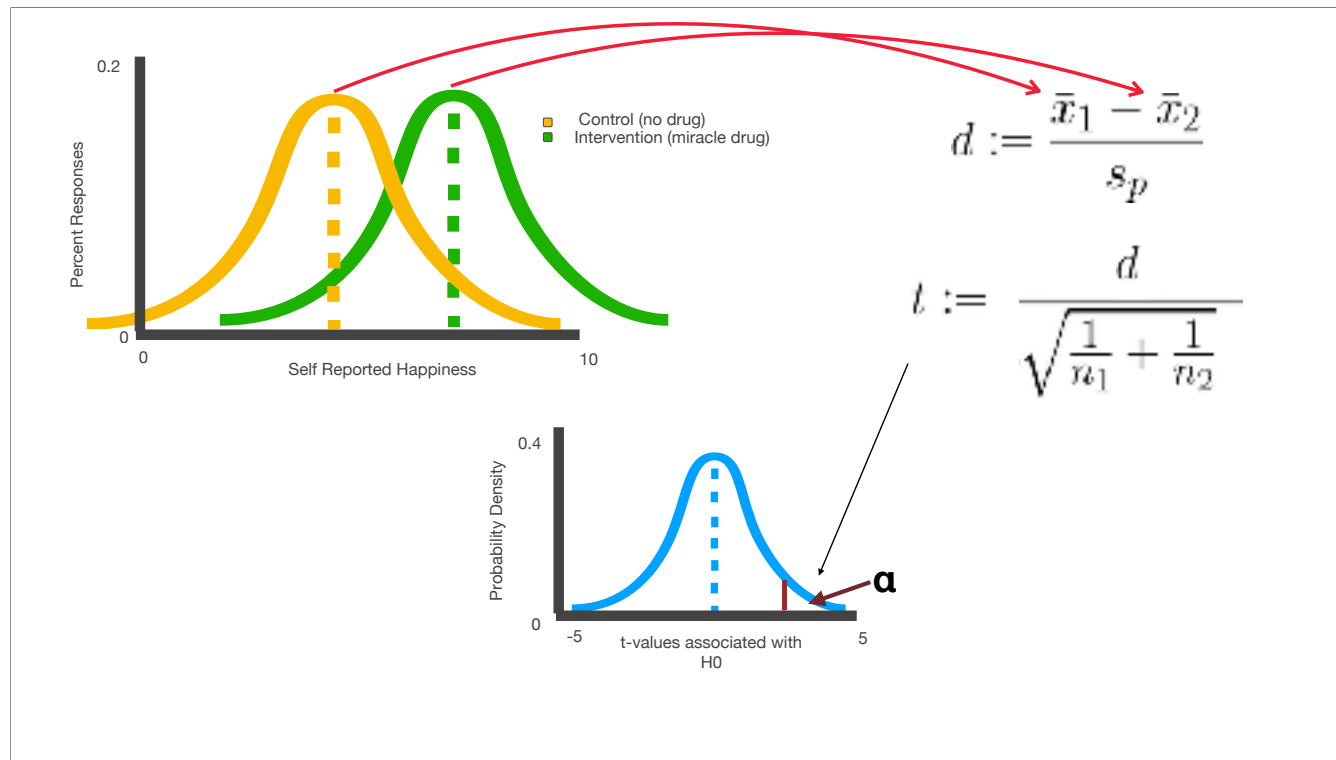
we care about *normalized* effect size too.

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen



so you probably remember the t-test from your basic stats class— our goal is to compare two groups that are experiencing an intervention, and see whether they have a difference large enough to allow us to *reject the null hypothesis*— i.e., the data between the groups is different enough that it seems very unlikely that we're seeing it simply as a result of random chance when sampling from the population.

Step 1: calculate normalized effect size. called hedges G or cohen's D — slightly different way of pooling the standard deviation, hedges G is more accurate for really small n(<20)

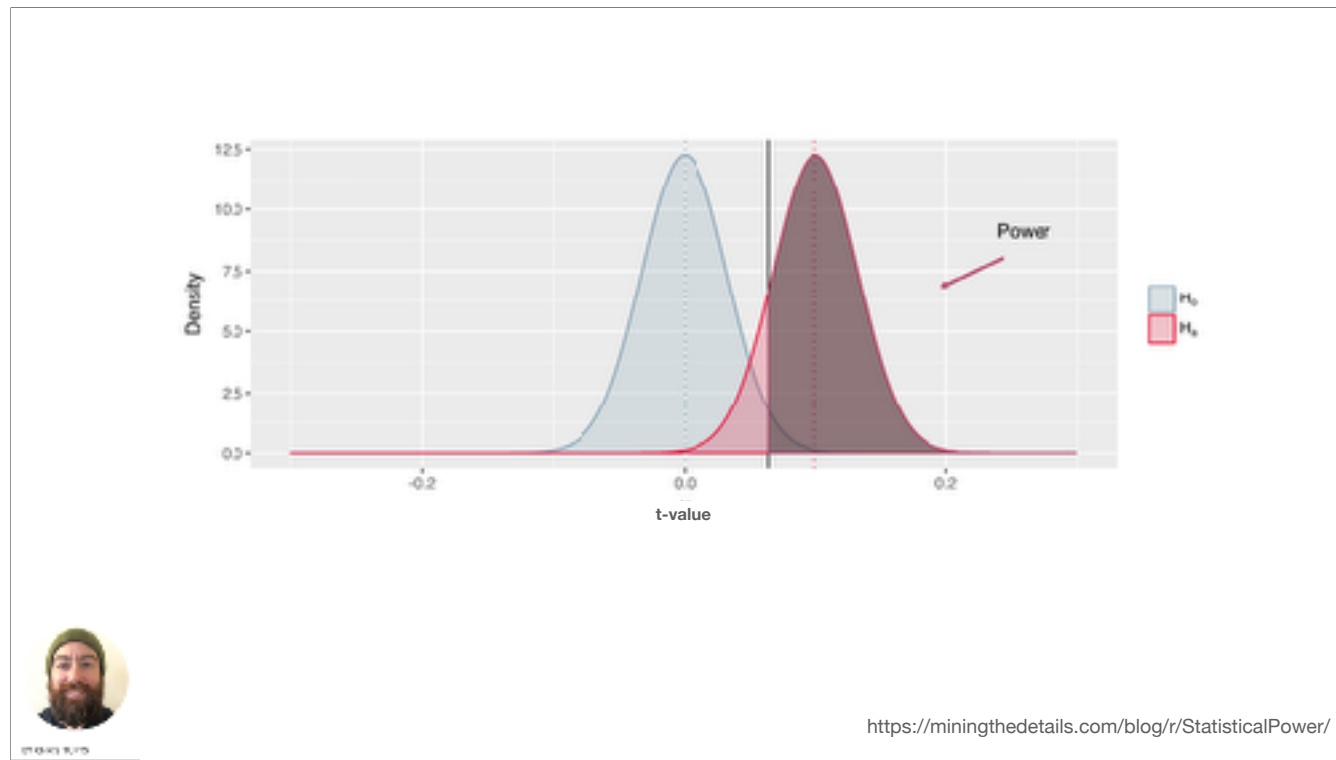
Step 2: Test statistic T = normalized effect size / sampling error

nice single number that gets bigger with big effects, smaller with high variance, bigger with big samples. Tells us how much we can trust that there's a real difference between these two.

we expect some small variations in effect size *even if both are drawn from the same underlying distribution in yellow* just from randomness, so in the case that our intervention and control data come from the same underlying data, we'd expect a distribution over probable t-values. That's what you see plotted here— the probability distribution of t-values if the null hypothesis is true— that there is no difference in the distributions underlying our data.

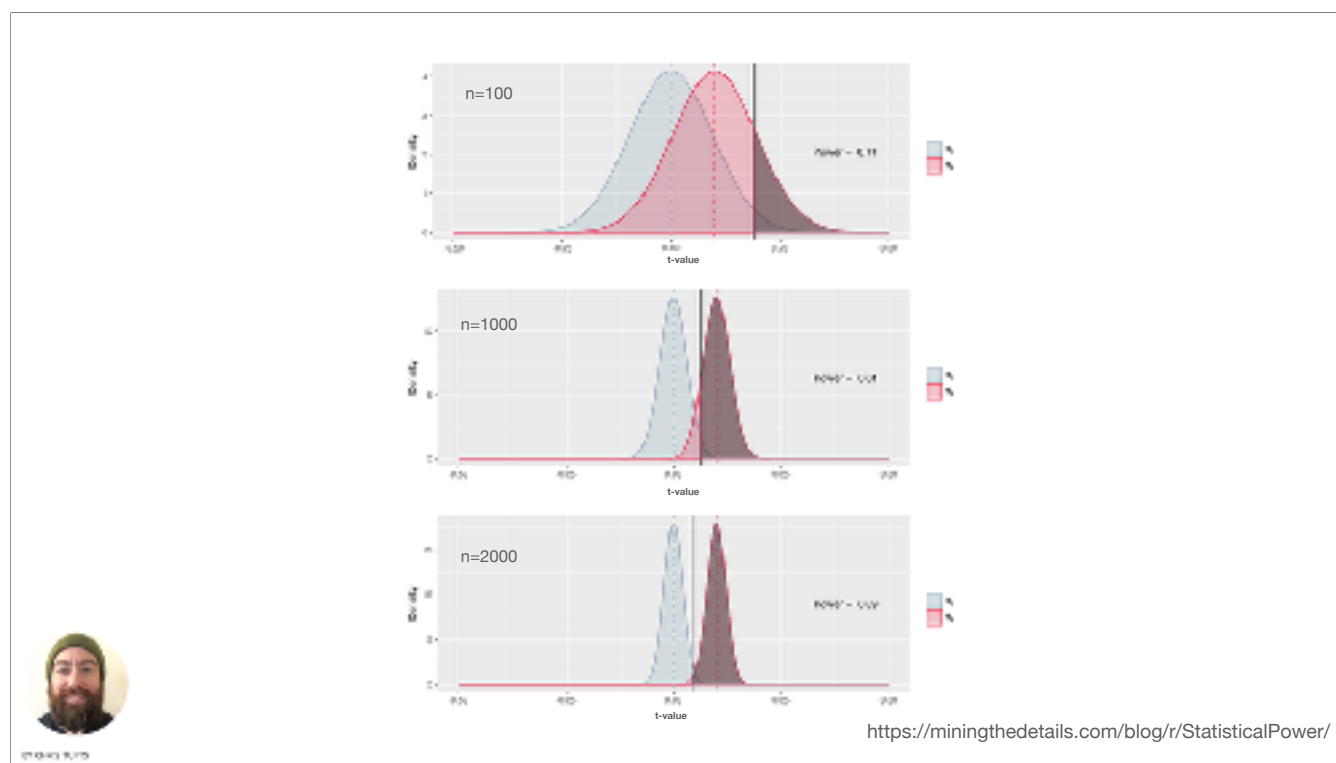
Again, we plot the distribution of test statistics we would see if there is no real difference between the groups— obviously if we sample from the yellow a few times, eventually we might accidentally draw a few left and a few right and it might look like an effect, and the t would be larger, but that's unlikely. With hypothesis testing we're always plotting the distribution of the test statistics we'd expect to see if there is no real underlying difference, and then trying to see if how unlikely it is that a t-value we get in our real test came from this distribution, so we can reject the idea that the data we're seeing is explained by noise in our sampling.

That's the p value; equivalently, it's the area under this probability density function for our t-statistic. We set a cutoff alpha (usually 0.05) and if p represents a less likely draw from this distribution we say we can 'reject the null hypothesis'. P-value represents how likely are we to see the data we have if we assume there is no difference between conditions; i.e. if we drew both yellow data and the green data from the same underlying distribution and they looked separated as an artifact of sampling.



We can also plot the distribution of t-values we'd expect for a hypothesis of ours, if we make a prediction about the real effect size and calculate our sample size (remember, t is determined by effect size and sampling error). Here we have the PDF of test-statistics we'd expect when there's no difference in blue (the null hypothesis), and the PDF of the test statistics in red we'd expect for our predicted effect size. *We have to make a guess about the real effect we expect to see to draw this plot.*

Once we do, we can reason about how likely we are to correctly accept an alternative hypothesis given our alpha cutoff from before. The more area of our PDF to the right of the cutoff, the more likely we get it right. In other words, if 80% of the red curve falls about the $\alpha=0.05$ line for the blue curve, we'd expect to correctly reject the null hypothesis 80% of the time, and fail to reject the null hypothesis (even though this effect, at this size, really does exist) the other 20% of the time. This is study power. You don't want to run low power studies— your study's odds of success and trustworthiness are defined by this concept!



this is something you should do for every study you run.
 Guess the effect size based on the literature, or a lower bound; calculate.

Here we see the differences in distributions as we increase our sampling size for the same true effect size (remember, $t = \text{effect size} / \text{sampling error}$). Notice how we get more and more likely to be able to separate the blue and red conditions (the null and alternative hypothesis) based on t-values when we increase the study power. To increase power at a given alpha, you can only increase sample size (or go study something else with a larger effect size).

I grabbed these nice plots from Chris Tufts, please check out his website!

Summary

Effect size (d) is what matters, and what you should look for. The most important quantity in the statistical test.

Normalized effect size (Cohen's D) tells us how big that change is relative to the original standard deviation. Both are important.

test statistics = normalized effect size / sampling error, and what we look at to determine whether we can reject the null hypothesis.

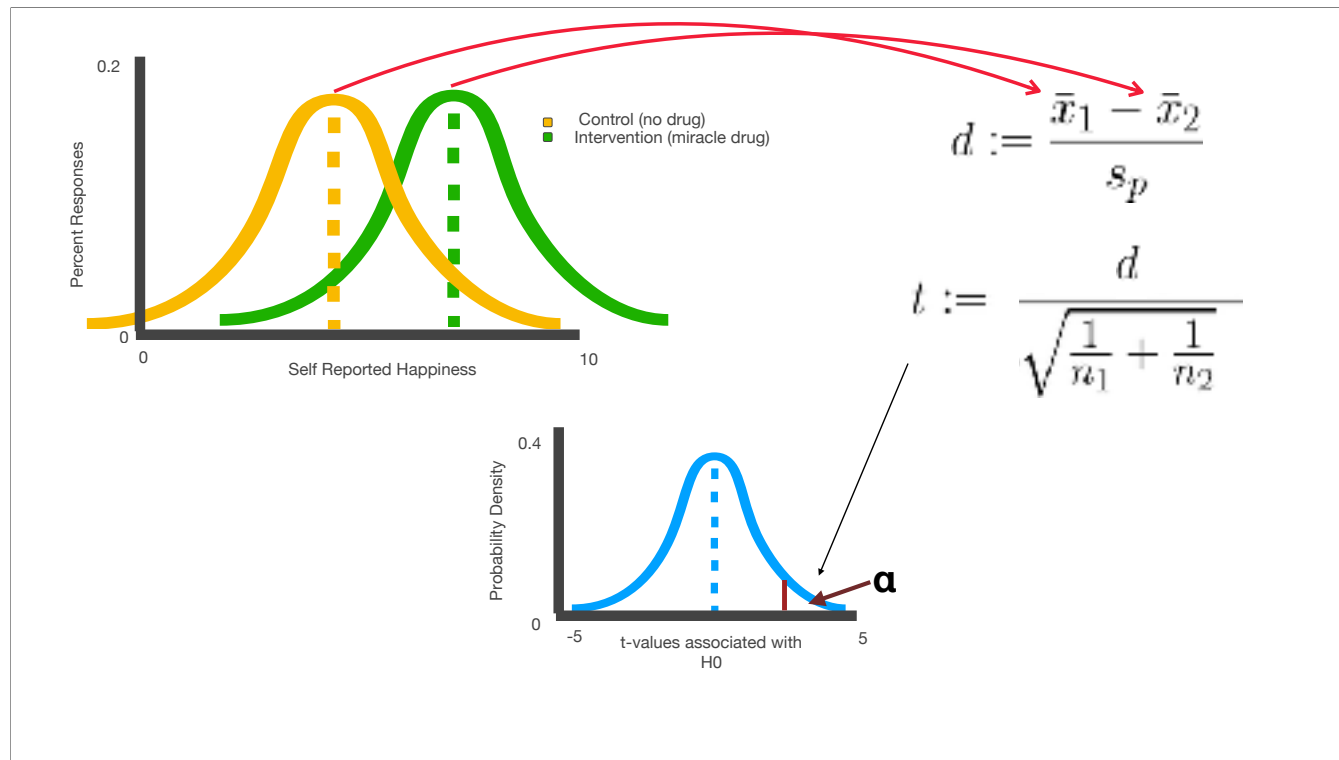
α is the 'significance threshold' or the likelihood of a False Positive *if we never test a true claim* (a.k.a. Type I error— we believe an effect is there when there isn't one). *Really* not good. Usually set at 0.05 and what we compare our p-value against to suggest we can reject the null hypothesis.

β is the likelihood of False Negative *if we always test a true claim of the specified effect size* (a.k.a. Type II error— we believe there isn't an effect when there is one). Not good, but not as bad as Type I. To mitigate this we need our study to have sufficient:

Power = (1- β) is our measure of how likely we are to detect an effect if there is an effect. Usually we set this to 80% before we consider doing a study worthwhile.

Effect size is what we actually care about, *much more than p-values*.

alpha/p-values — the probability of seeing the data you're seeing because of random chance if they were drawn from the same underlying distribution.



so what is wrong with this picture??? Where does it go wrong??

p-values tell us something about how likely we are to have seen our data if there was no difference; how common the t-value we calculate is if all the data was drawn from the same distribution.

ALPHA!!!

Categorical thinking is the cardinal sin of modern statistical practice.

86% of statistics teachers and
40% of psychology professors get this wrong.

Gigerenzer, Gerd. "Mindless statistics." *The Journal of Socio-Economics* 33.5 (2004): 587-606.

Gigerenzer, Gerd. "Statistical rituals: The replication delusion and how we got there." *Advances in Methods and Practices in Psychological Science* 1.2 (2018): 198-218.

Professor Alice

testing very unlikely hypotheses
big, life-changing results

1/1000 are real effects

1 experiment at 90% power

Professor Bob

quantifying effect size of common sense hypotheses
confirmation and insight from unsurprising results

4/5 are real effects

1 experiment at 90% power

Professor Alice

testing very unlikely hypotheses
big, life-changing results

1/1000 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.001 * power (0.9) = 0.0009

Professor Bob

quantifying effect size of common sense hypotheses
confirmation and insight from unsurprising results

4/5 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.8 * power (0.9) = 0.72

Professor Alice

testing very unlikely hypotheses
big, life-changing results

1/1000 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.001 * power (0.9) = 0.0009

likelihood of p value < 0.05 being a real effect?

$0.05 * 0.0009 = 0.000045$ chance that they co-occur.
 $0.0009 / (0.0009 + (0.05 - 0.000045)) = 1.8\%$

Professor Bob

quantifying effect size of common sense hypotheses
confirmation and insight from unsurprising results

4/5 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.8 * power (0.9) = 0.72

likelihood of p value < 0.05 being a real effect?

$0.05 * 0.72 = 0.036$ chance that they co-occur.
 $0.72 / (0.72 + (0.05 - 0.036)) = 98.1\%$

you might have an instinct. That I've just said Bob does good work and Alice does bad work, because Bob's work is trustworthy at $p=0.05$ and Alice's work is untrustworthy at $p=0.05$. I have that instinct. That's exactly the instinct we need to destroy because it's wrong. We need both Alice and Bob. Most of us are closer to Alice than to Bob.

I want to emphasize what I've not just said. I said **for the same p-value, you should have more trust in Bob's work**. If that makes you feel emotionally like I've said Bob is a better researcher, or Alice's work is worse or less valuable or less trustworthy, it's because you have so deeply accepted the incorrect way of interpreting p-values that it has infiltrated your subconscious.

I already said that Alice and Bob are good researchers. They're pre-registering their high powered studies, they're ethical and following the right practices, and I would rather be Alice (and I suspect most of you would too).

This is the calculation you need to be doing in your head. And this betrays the biggest point of confusion for people about what the p-value means, and how to interpret it.

The calculation we just did is call the PPV, and it's what we actually care about. How likely is this to be true? How likely is this to be a real effect? And as we just saw, it depends on our prior beliefs— Alice was testing whether your mother has psychic powers, Bob was testing whether kids like candy. If they both come back with $p \leq 0.05$, you should interpret that information differently.

Professor Alice

testing very unlikely hypotheses
big, life-changing results

1/1000 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.001 * power (0.9) = 0.0009

likelihood of p value < 0.05 being a real effect?

$0.05 * 0.0009 = 0.000045$ chance that they co-occur.
 $0.0009 / (0.0009 + (0.05 - 0.000045)) = \mathbf{1.8\%}$

odds a positive is a False Positive (Type I Error) = 98.2%
odds of a False Positive (Type I Error) in general = 0.05 - 0.000045 = 4.9955%

Professor Bob

quantifying effect size of common sense hypotheses
confirmation and insight from unsurprising results

4/5 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.8 * power (0.9) = 0.72

likelihood of p value < 0.05 being a real effect?

$0.05 * 0.72 = 0.036$ chance that they co-occur.
 $0.72 / (0.72 + (0.05 - 0.036)) = \mathbf{98.1\%}$

odds a positive is a False Positive (Type I Error) = 1.9%
odds of a False Positive (Type I Error) in general = 0.05 - 0.036 = 1.4%

Professor Alice

testing very unlikely hypotheses
big, life-changing results

1/1000 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.001 * power (0.9) = 0.0009

likelihood of p value < 0.05 being a real effect?

0.05*0.0009 = 0.000045 chance that they co-occur.
 $0.0009 / (0.0009 + (0.05-0.000045)) = 1.8\%$

odds a positive is a False Positive (Type I Error) = 98.2%
odds of a False Positive (Type I Error) in general = 0.05 - 0.000045 = 4.9955%

Professor Bob

quantifying effect size of common sense hypotheses
confirmation and insight from unsurprising results

4/5 are real effects

1 experiment at 90% power

odds of 1 p-value < 0.05 from chance = 0.05 (alpha)

odds of real effect leading to p-value < 0.05 = 0.8 * power (0.9) = 0.72

likelihood of p value < 0.05 being a real effect?

0.05*0.72 = 0.036 chance that they co-occur.
 $0.72 / (0.72 + (0.05-0.036)) = 98.1\%$

odds a positive is a False Positive (Type I Error) = 1.9%
odds of a False Positive (Type I Error) in general = 0.05 - 0.036 = 1.4%

for $p < 0.05$, we expect 5% false positive results *if we never test a real effect*.

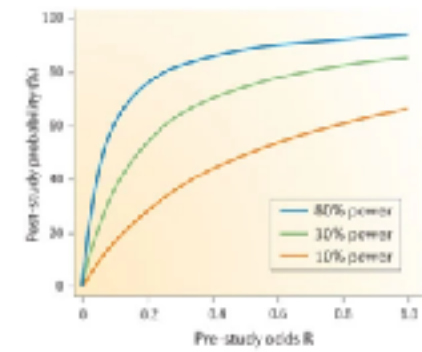
The likelihood that an effect is real, given a certain p-value, *depends on your prior*.

a p-value of 0.05 means something very different for Alice and Bob and should be interpreted differently.

Positive Predictive Value (PPV)

$$PPV = (1-\beta)*R / ((1-\beta)*R + \alpha)$$

Figure 4: Positive predictive value as a function of the pre-study odds of association for different levels of statistical power.



Nature Reviews | Neuroscience

Button, K., Ioannidis, J., Mokrysz, C. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14, 365–376 (2013). <https://doi.org/10.1038/nrn3475>

here R = 'odds ratio'. (for Bob in our example above, it's 4/1, for Alice it's 1/999). This is a mathematically precise definition of what we started to walk through before.

No scientific worker has a fixed level of significance *at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.*

- Fisher (father of statistics)

DO NOT CONFUSE PPV WITH P!!!!

Does a p value < 0.05 mean the effect is 95% likely a real effect?

Does a p value < 0.05 mean the effect is 95% likely a real effect?

NO, that is the PPV. ('inverse probability error')

Does a p value < 0.05 mean the effect is 95% likely a real effect?

NO, that is the PPV. ('inverse probability error')

Does a p value < 0.05 mean the same exact experiment will find a 'significant result' effect 95% of times you run it?

Does a p value < 0.05 mean the effect is 95% likely a real effect?

NO, that is the PPV. ('inverse probability error')

Does a p value < 0.05 mean the same exact experiment will find a 'significant result' effect 95% of times you run it?

NO, that's the study power. ('the replication fallacy')

Does a p value < 0.05 mean the effect is 95% likely a real effect?

NO, that is the PPV. ('inverse probability error')

Does a p value < 0.05 mean the same exact experiment will find a 'significant result' effect 95% of times you run it?

NO, that's the study power. ('the replication fallacy')

Does a hypothesis test tell you anything certain about a hypothesis being true or false?

Does a p value < 0.05 mean the effect is 95% likely a real effect?

NO, that is the PPV. ('inverse probability error')

Does a p value < 0.05 mean the same exact experiment will find a 'significant result' effect 95% of times you run it?

NO, that's the study power. ('the replication fallacy')

Does a hypothesis test tell you anything certain about a hypothesis being true or false?

NO. ('illusion of certainty')

THE WORLD IS CHANGING

Editorial
The ASA Statement on p -Values: Context, Process, and Purpose
Ronald L. Wasserstein & Nicole A. Lazar
Pages 120-133 | Accepted author version posted online 07 Mar 2016; Published online 09 Jun 2016

Editorial
Moving to a World Beyond “ $p < 0.05$ ”
Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar
Pages 1-19 | Published online 20 Mar 2019

It's time to talk about ditching statistical significance

Scientists rise up against statistical significance

the board of the ASA tasked Wasserstein to assemble a panel of experts.

Reproducibility project re-tested 100 studies from top 3 psych journals, <50% gave 'statistically significant results'.

THE WORLD IS CHANGING

Editorial

The ASA Statement on p -Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Pages 729–732 | Accepted author version posted online 01 April 2016; Published online 20 April 2016

Editorial

Moving to a World Beyond “ $p < 0.05$ ”

Ronald L. Wasserstein, Allen L. Schimm & Nicole A. Lazar

Pages 7–12 | Published online 20 March 2016

nature

It's time to talk about ditching statistical significance

nature

Scientists rise up against statistical significance

THE WORLD IS

In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and p -values. The issue also included many commentaries on the subject. This month's special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on statistical inference in the 21st century, a world beyond $P < 0.05$. The editors introduce the collection with the caution "Do not say 'statistically significant'". Another article with dozens of signatures also calls on authors and journal editors to disuse those terms.

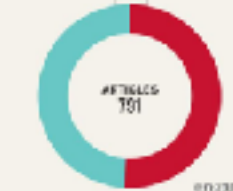
We agree, and call for the entire concept of statistical significance to be abandoned.

WRONG INTERPRETATIONS

An analysis of 761 articles across 8 journals found that around half statistically assume non-significance means no effect.

Accurately interpreted
49%

Wrongly interpreted
51%



*Data taken from F. Eklake et al., *Journal of Management Science*, 17(4), 2019, pp. 1-11. For more on this, see *Statistical Science*, 34(2), 2019, pp. 1-11. DOI: 10.1214/18-STATS104. © 2019 by the Institute of Mathematical Statistics. Published online by Cambridge University Press.

Source: www.elsevier.com/locate/jms

Do not say 'statistically significant' ever again.

report full p-values, not " $p < 0.05$ ".

Be suspicious of people who do those things.

Interpret p-values cautiously based on your priors.

What does that mean about 'replication'?

ERR vs. 'colloquial' replication

replication is a categorical concept!! it depends on 'statistical significance'!!

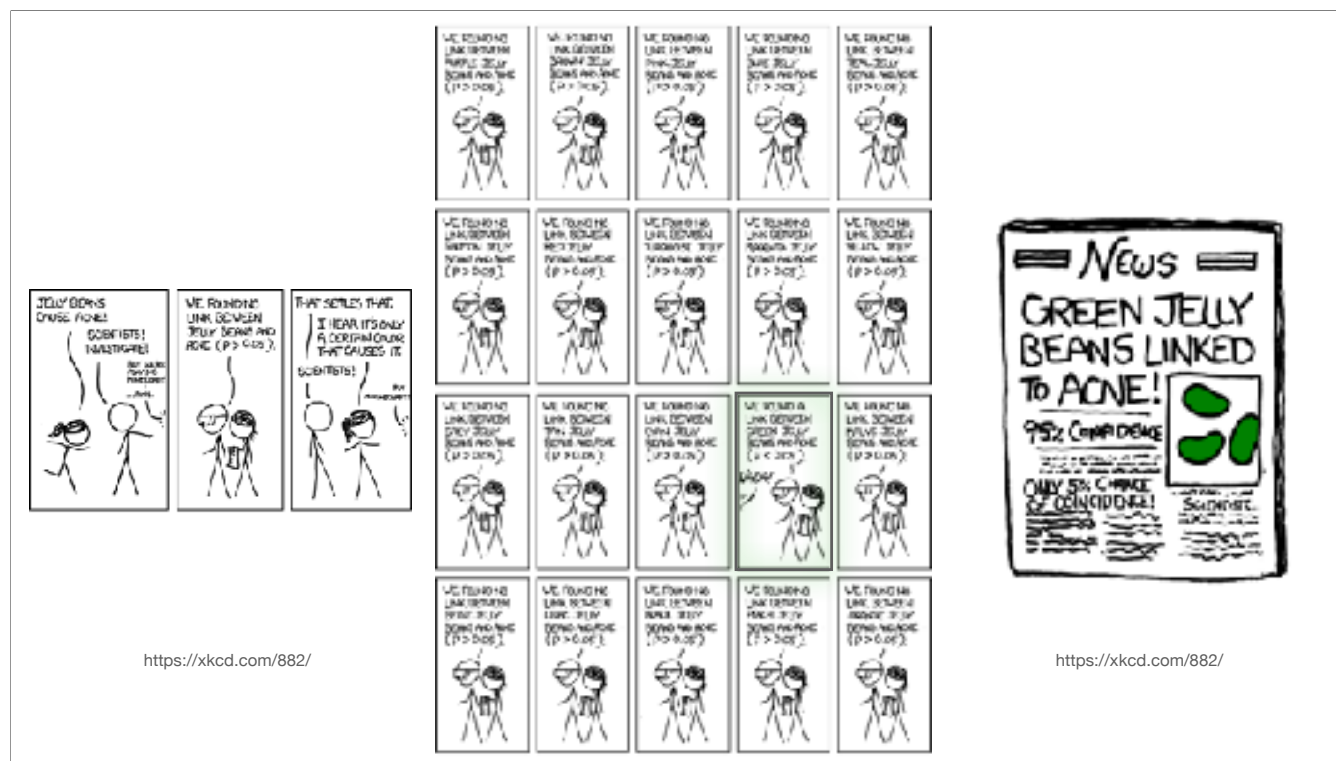
not great, and a heuristic.

ERR (how likely is a study to give the same result if done exactly the same way, i.e. an estimate of study power)

'replication' (study done at much higher power to see if you can reject the null and to get a much more accurate sense of the effect size)

Categorical thinking is the cardinal sin of modern statistical practice.

end rant on p-values.



so what happens when we apply statistical significance in scientific inquiry? Keep this XKCD in mind.

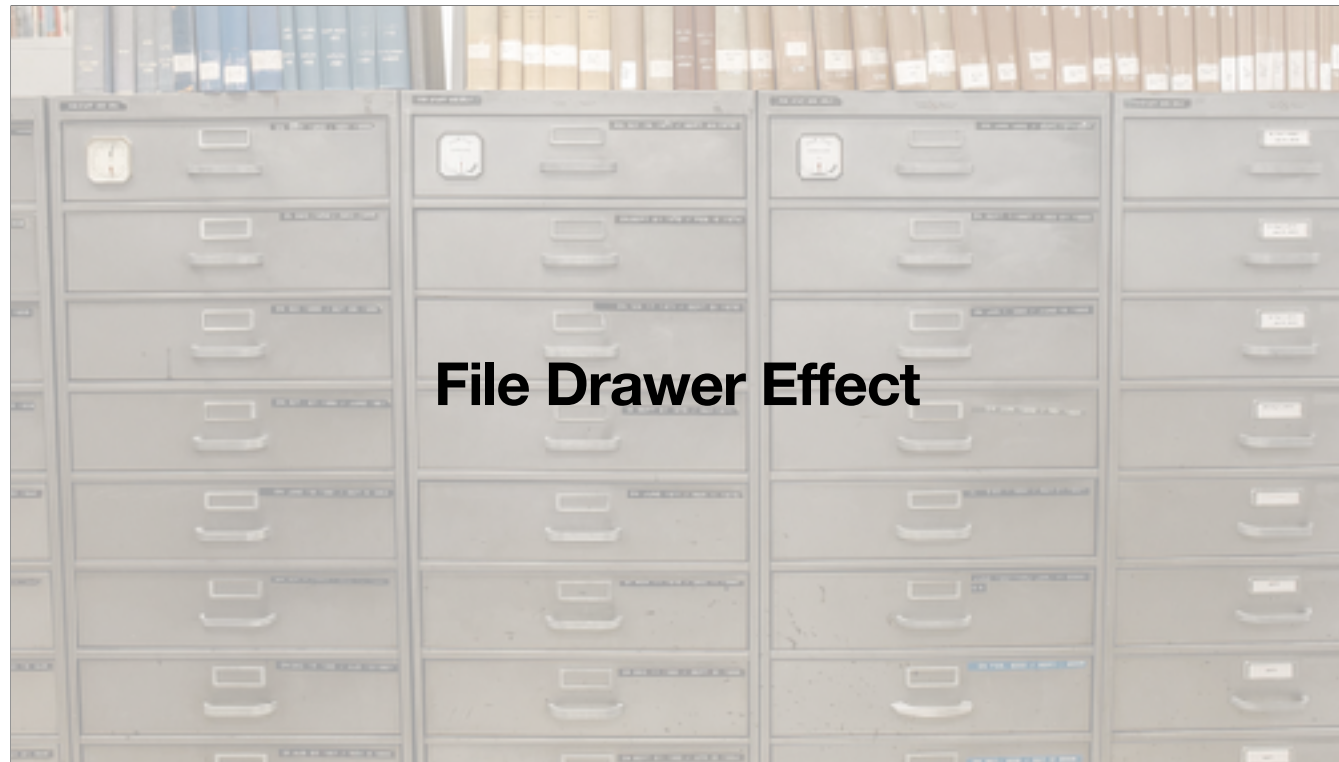
point 1. We have to do the middle bit. Test lots of things. This is only a problem when we *don't* see what the scientist did in complete detail. If we can reason about how many hypotheses were tested, or how many times people ran similar experiments, we can easily contextualize the results and their *p-values*.

point 2. now we're here reading this newspaper, with no insight into what happened. looking at *how we got where we are*, we can learn what to look out for in the research literature to (1) tip us off and reason backwards about what happened, reconstruct the process, or (2) identify when it's clearly gone wrong and ignore it.



journals allowing only 'interesting results' in. Journals will do the job that XKCD.

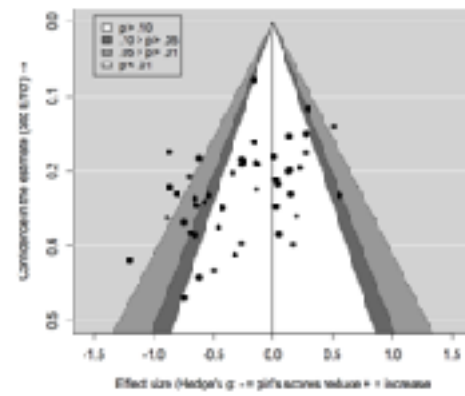
20 submissions. 1 is published.



self-censorship by researcher.

don't submit a paper because it will be rejected. 20 try, 1 succeeds, winner publishes, everyone else doesn't try.

Publication Bias and File Drawer Effects



Meta-analysis of [stereotype threat](#) on girls' math scores showing asymmetry typical of publication bias. From Flore, P. C., & Wicherts, J. M. (2015)

Rank	Journal	2019	2018	2017	17-19	14-16	10-13	Change
1	Journal of Religion and Health	85	72	74	72	51	75	3
2	Journal of Individual Differences	86	65	55	60	55	76	4
3	Journal of Business and Psychology	87	73	66	83	76	77	5
4	Journal of Research in Personality	87	78	75	63	75	72	+9
5	Journal of Happiness Studies	85	66	57	70	79	69	-13
6	Journal of Occupational Health Psychology	88	66	66	68	72	68	2
7	Journal of Youth and Adolescence	85	66	72	74	50	75	1
8	Journal of Homosexuality	84	75	53	64	70	68	-15
9	Journal of Research on Adolescence	84	64	55	71	57	73	-2

<https://replicationindex.com/2020/02/21/replicability-rankings-of-120-psychology-journals-2010-2019/>

distribution of p-values/effect sizes.

funnel plot — ‘high precision’ studies should have an effect size that is in the middle. Skew is evidence of bias. (Standard error or sample size).

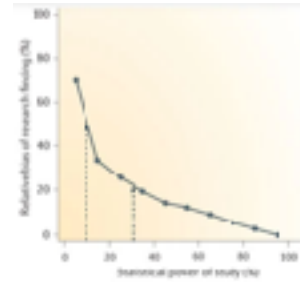
120 top journals, auto-analyzed by Ulrich Shimmack.

Publication Bias and File Drawer Effects

Beware of 'in vogue' research topics.

Publish all of your findings (open access).

The Winner's Curse



Nature Reviews | Neuroscience

The winner's curse refers to the phenomenon that studies that find evidence of an effect often provide inflated estimates of the size of that effect. Such inflation is expected when an effect has to pass a certain threshold – such as reaching statistical significance – in order for it to have been 'discovered'. Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large. If, for example, the true effect is medium-sized, only those small studies that, by chance, estimate the effect to be large will pass the threshold for discovery (that is, the threshold for statistical significance, which is typically set at $p < 0.05$). In practice, this means that research findings of small studies are biased in favour of inflated effects. In contrast, large, high-powered studies can readily detect both small and large effects and so are less biased, as both over- and underestimation of the true effect size will pass the threshold for 'discovery'. We optimistically

Button, K., Ioannidis, J., Mokrysz, C. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14, 365–376 (2013). <https://doi.org/10.1038/nrn3475>

'Edlin Factor'

Assume the real effect size is ~1/2 to 1/100 of what is reported.

<https://statmodeling.stat.columbia.edu/2014/02/24/edlins-rule-routinely-scaling-published-estimates/>

corollary. We are 'rolling the die' a bunch of times to get a p -value < 0.05 . If our studies are underpowered, we will get hugely inflated effect sizes. The test statistic = effect size / sampling error. Big sampling error means big effect size required to hit that value.

Heuristics.



those are ways we systemically have this problem. Now let's look at how an individual researcher might effectively accomplish the same goal— test a ton of hypotheses and only report on what's hitting.

how *NOT* to be a bad researcher

this section is called 'how *NOT* to be a bad researcher'

QRPs or 'Researcher Degrees of Freedom'

- **'Optimal Stopping'** (high variance, small sample studies)— *beware research with unusual Ns*
- **Combining Disparate Subgroups/Tests Together**— *beware research that combines multiple 'sub-experiments' into one and only reports analysis on that result*
- **Testing many hypotheses and only reporting the successes**— *beware research with odd or specific mediating variables, or other unusual caveats*
- **Running many tests to prove a point, and only reporting the successes**— *beware 'motivated researchers' that clearly have a narrative agenda, and whose research all cumulatively backs up a given worldview or theory; beware papers without a clearly articulated a priori hypothesis and many statistical results*



"I'm all for rigor, but I prefer other people do it. I see its importance—it's fun for some people—but I don't have the patience for it. If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?'"

- Daryl Bem, in Engber, 2017



it's okay to have multiple hypotheses; *as long as we know what you did*. You should revise your internal alpha down.

Report everything, including what didn't hit.

What's NOT okay is to *create the hypotheses after seeing the data*, if we act like we had those hypotheses originally.

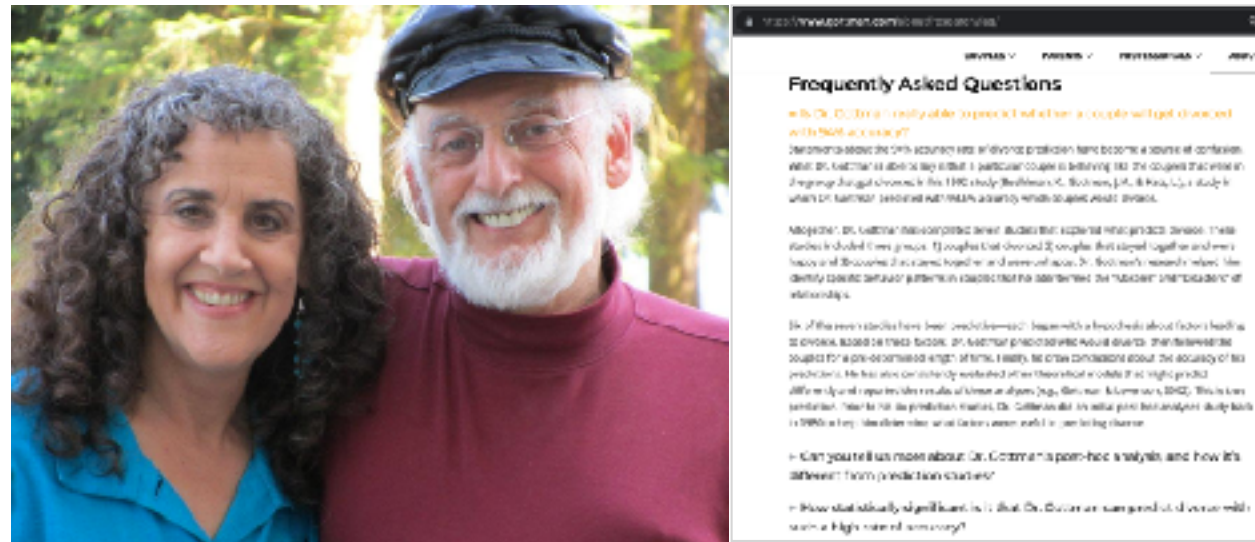
HARKing

- **Hypothesizing After the Results are Known**
- ‘data dredging’, ‘fishing’, ‘cherry-picking’, ‘mining’, ‘p-hacking’
- try tons of hypotheses until something gives you $p < 0.05$; report that as though it was your original hypothesis.
- try a ton of slight variations of modeling, statistical techniques, and data preprocessing; report the ones that give you $p < 0.05$.
- look at your data; *come up with a likely explanation, decide on a method for cleaning it, or pick a method to fit a model to the data after seeing it.*

there are lots of potential relationships that might appear in your data. infinite number of potential relationships. If you predict a curve, and that curve then appears, that's *really powerful evidence*. That's the expectation for how science works. If you instead look for a relationship, and see a line or a curve or an exponential appear somewhere, and *then fit a line to it and act like you had predicted it*, that's incredibly misleading. *Some* relationship will appear.

So widespread. I think every researcher, unless they have been very, very well-trained, has done this.

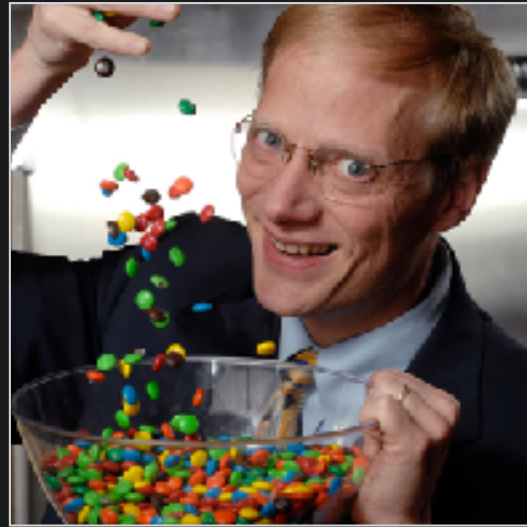
Exploratory vs Confirmatory Studies



what if you really don't know how to treat your data? Or you have no idea what you're doing? Or you're using your study to develop a hypothesis? That's fine, too, but we need to be explicit about it. That's called an exploratory study, and you must be very very clear that it's what you're doing, and draw no conclusions (only state new hypotheses).

It's okay to try a ton of things on the data if we explicitly call out that we're exploring with the data and developing theories, and we report what we did and how, completely. We don't make any causal claims or suggest the data proves something.

Example is Gottman; conflated exploratory and confirmatory studies. 20 couples that were heavily measured, wait to see who divorces, now we have a group of 10 couples that divorced and 10 couples that didn't. Look to see what is different about divorced 10 from non-divorced 10. Turns out, it's possible to separate these couples based on some aspects of their relationship with 94% accuracy. But you could separate *any* two groups of 10 couples based on certain features of their relationships; couples vary a lot. This is exploratory; we can now look and say 'do any of these things make sense' to explain divorce, and use that to predict brand new couples will divorce, and see if it works, in a confirmatory study.



"P-hacking and MTurk-iterating isn't helpful to science, and it's one of the reasons our lab seldom cites on-line studies. However, P-hacking shouldn't be confused with deep data dives – with figuring out why our results don't look as perfect as we want.

With field studies, hypotheses usually don't "come out" on the first data run. But instead of dropping the study, a person contributes more to science by figuring out when the hypo worked and when it didn't.

a tale of two young researchers

...When she arrived, I gave her a data set of a self-funded, failed study which had null results (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people 1/2 as much as others). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses.

Eventually we started discovering solutions that held up regardless of how we pressure-tested them."

- Brian Wansink, Blog post 1/17/2016

Researcher Analysis



The screenshot shows a table with 8 columns: RANK, NAME, TESTS, CER, EDR, ERR, FDR, and ALPHA. The table is sorted by RANK in ascending order. The first five rows are visible, showing researchers and their corresponding metrics. The table has a search bar at the top right and a dropdown menu for the number of entries shown (currently set to 10).

RANK	NAME	TESTS	CER	EDR	ERR	FDR	ALPHA
1	Robert A. Cyrono	53	87	60	90	1	.85
2	Allison L. Skinner	227	59	66	85	1	.83
3	David Marcusovits	708	83	79	85	1	.85
4	Linda J. Skiba	522	68	75	82	1	.86
5	Todd K. Shackelford	309	77	73	82	2	.83

<https://replicationindex.com/2021/01/19/personalized-p-values/>

reduce alpha from 0.05 to 0.01

also they could be shooting for a needle in a haystack vs testing obvious things, as we discussed earlier.

Beware, incredibly noisy, preliminary analysis
different alpha values for each.

reduce alpha 0.05 to 0.01 as a heuristic!!

Meta-statistical Tools for Bias Detection

[stat.io](#) and [pubpeer.com](#)

funnel plots, hodge's g forest plots

p-curve, z-curve, meta-analysis of observed powers

Carlisle-Stouffer-Fisher Method (<https://pubmed.ncbi.nlm.nih.gov/28786843/>)

GRIM - Granularity-Related Inconsistency of Means (<https://peerj.com/preprints/2064.pdf>)

TIVA - Test of Insufficient Variance (<https://replicationindex.com/2014/12/30/tiva/>)

CORVIDS - Complete Recovery of Values in Diophantine Systems (<https://github.com/katherinemwood/corvids>)

SPRITE - Sample Parameter Reconstruction via Iterative Techniques <https://peerj.com/preprints/26968v1/>)

lots of meta-statistical tools to help us. Check blogs, Noah will talk about next class.

Always perform an a priori power analysis

Make an educated guess about your hypothesized effect size, and see how large of an N you need to detect it 80% of the time.
Don't waste your time on studies that are unlikely to reveal an effect (and would be untrustworthy if they did).

<https://emj.bmj.com/content/emered/20/5/453.full.pdf>

<https://stats.oarc.ucla.edu/other/gpower/> (UCLA's G*Power tool)

<https://clincalc.com/stats/samplesize.aspx> (online calculator)

STATISTICS

An introduction to power and sample size estimation

S R Jones, S Cooley, M Harrison

The importance of power and sample size estimation for study design and analysis.

OBJECTIVE

1. Explain the importance of power and sample size estimation in clinical research and analysis.

2. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

3. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

4. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

5. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

6. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

7. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

8. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

9. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

10. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

11. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

12. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

13. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

14. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

15. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

16. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

17. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

18. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

19. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

20. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

21. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

22. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

23. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

24. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

25. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

26. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

27. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

28. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

29. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

30. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

31. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

32. Understand the relationship between sample size, confidence intervals, and diagnostic tests.

Sample Size Calculator

Determines the minimum adequate study power

UCLA's G*Power tool

Online calculator

Study Group Design

The relationship between study groups

The study groups will each receive different treatments.



Always do a power analysis! resources here. We talked about the danger of underpowered research.

Beware 'n=20' research.

That researcher *probably didn't do an a priori power analysis*, so you should do one before you read their conclusions.

Do you think their study was sufficiently powered to catch a result of an intuitive size?

and a corollary, beware papers where it seems likely the researcher didn't do a power analysis. You should do one— how big of an effect do you think is likely for the study they're doing? How large of an N do they need? If it's far off, the study is untrustworthy.

A Priori Power and Pre-registration

- Do your a priori power analysis and don't run under-powered studies!
- OSF Preregistration: <https://osf.io/prereg/>
- UPenn Credibility Lab Preregistration: <https://aspredicted.org/>



<https://www.cos.io/initiatives/prereg>

link at bottom has good resources— which journals accept preregistration and will give papers a badge, which you should look for!

how to participate and get journals on board.

I. METHODS

MISAPPLIED STATISTICS

P-Hacking, Publication Bias, File Drawer Effects, HARKing, QRPs, Researcher DOF

II. SYSTEMS

SYSTEMIC INCENTIVES

Fraud, Hype, Motivated Misinterpretation, Overgeneralization, Narrative Support, Status Quo Bias

III. ONTOLOGIES

FUNDAMENTAL ASSUMPTIONS


Psychometrics, Taxometrics, Analytic/Gestalt, Idiographic/Nomothetic, Statistical/Causal Reasoning

Other Issues

Fraud and Hype

The Mind of a Con Man

f t + Read in app



Dieterik Stapel, a Dutch social psychologist, perpetrated an audacious academic fraud by making up studies that told the world what it wanted to hear about human nature. *Kiss Broken for The New York Times*

By Tuhijit Bhattacharjee
April 26, 2013

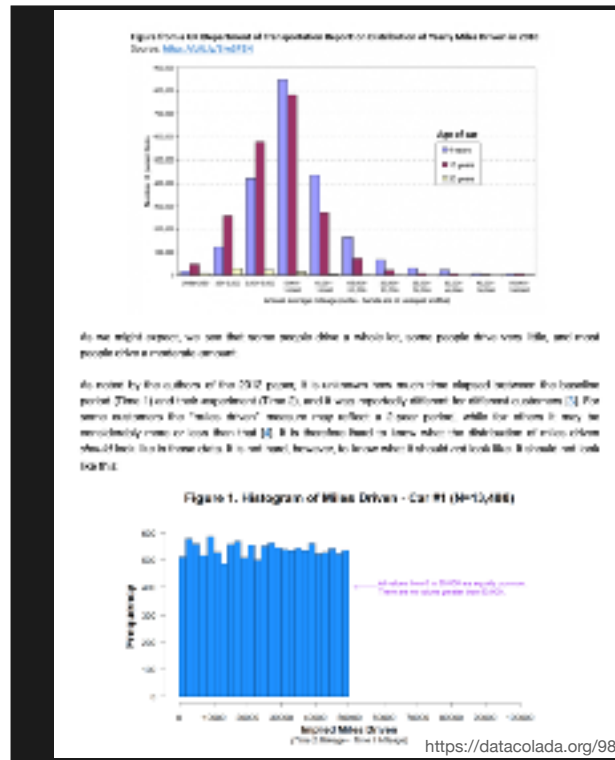
58 retractions

physical environments that are more disordered promote stereotyping and discrimination [in Science]

carnivores are more selfish than vegetarians

ads can affect whether and how consumers think about the self

Dirkje Stapel



THE TIMES OF ISRAEL

Behavioral researcher says he 'undoubtedly made a mistake' in false data scandal

Dan Ariely insists his actions were innocent, defends other controversial moves, believes his reputation will recover

By TCI STAFF
4 Sep 2021, 11:29 am | 98

Prof. Dan Ariely (Channel 12 screenshot)

Dan Ariely recently in the news, distribution of data obviously faked.

Retractions

<https://retractionwatch.com>

<https://retractiondatabase.org>

The screenshot shows the search interface for the Retraction Watch Database. At the top, it says "The Retraction Watch Database" and "Please see this [user guide](#) before you get started". The interface is divided into several sections:

- Author:** A text input field with a dropdown arrow.
- Year:** A text input field with a dropdown arrow.
- Journal:** A text input field with a dropdown arrow.
- Subject:** A text input field with a dropdown arrow.
- Country:** A text input field with a dropdown arrow.
- Source:** A text input field with a dropdown arrow.
- Publication:** A text input field with a dropdown arrow.
- Retraction:** A text input field with a dropdown arrow.
- Notes:** A text input field.
- DOI:** A text input field.

On the right side, there are two sections for filtering results:

- Original paper:** A dropdown menu with options "Yes" and "No".
- Retraction or Correction:** A dropdown menu with options "Yes" and "No".

At the bottom left, there is a "Clear Search" link. At the bottom right, there is a green "Search" button.

great central resource for retractions

May 21, 2021 | By Christine Clark

SHARE f w In   

A New Replication Crisis: Research that is Less Likely to be True is Cited More

Papers that cannot be replicated are cited 153 times more because their findings are interesting, according to a new UC San Diego study

Remarkably, only 12 percent of post-replication citations of non-replicable findings acknowledge the replication failure,

The largest gap was in papers published in Nature/Science: non-replicable papers were cited 300 times more than replicable ones.

Yearly citation counts reveal a pronounced gap between papers that replicated and those that did not. On average, papers that failed to replicate are cited 16 times more per year. This gap remains even after the replication project is published.

“Remarkably, only 12 percent of post-replication citations of non-replicable findings acknowledge the replication failure,” the authors write.

**Read the papers you cite beyond
the title and abstract.**

**Check for retractions and
replications.**

Hype

Ecosystem that rewards big, eye-catching claims

Abusing 'statistically significant' to mean 'significant'

Categorical thinking instead of continuous thinking

Overselling effects, Overgeneralizing results, and Narrative Cherry-picking

Misleading Plots

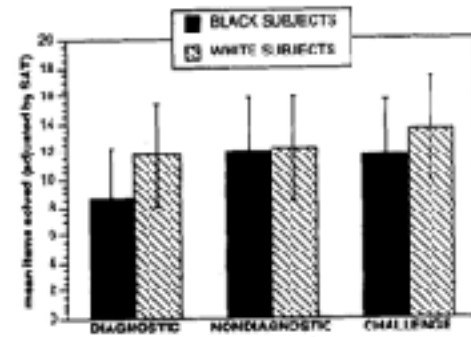
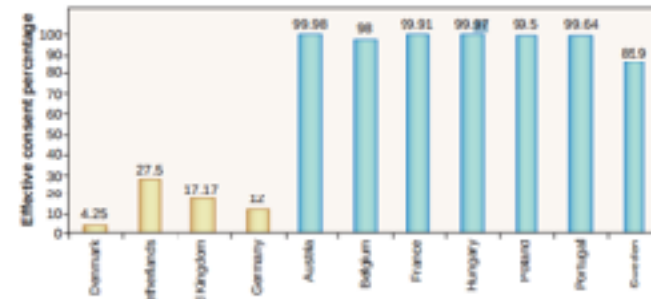


Figure 1. Mean test performance Study 1.

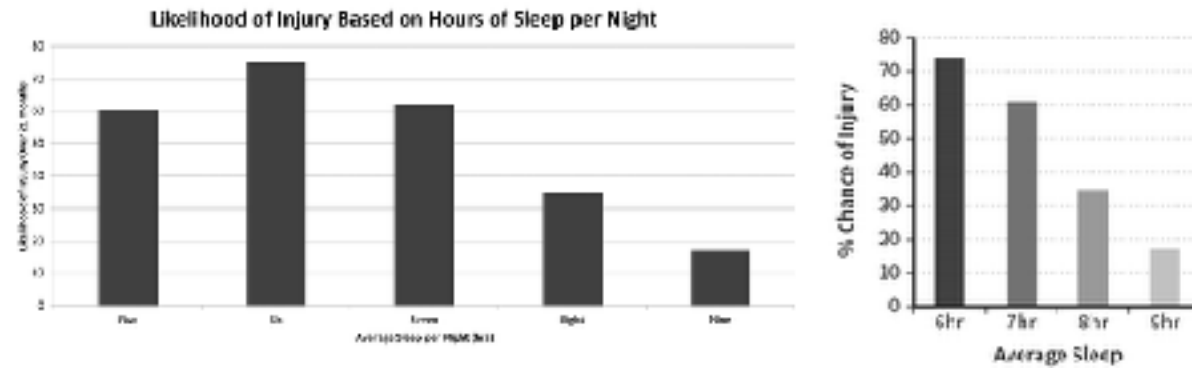


Effective consent rates, by country. Explicit consent (opt-in, gold) and presumed consent (opt-out, blue).

relative effect sizes. relative statistics. Famously contentious/misleading stereotype threat original paper.

misleading underlying assumptions. 'Default choice' work when really the blue is 'presumed consent' (citizens are never asked, no default on a form, have to actively initiate opting out), somehow distorted to imply that people are overwhelmed by complex decisions and go with defaults on a form for organ transplant.

Misleading Plots



Micheal Walker's 'Why We Sleep'. Real data on the left, data reported in the book on the right to make his point. Clear misconduct for rhetorical end.

Mind the Axes.
(Check what data is in the figure)

Check the Error Bars.
($1.96 * \text{Standard Error} = 95\% \text{ Confidence Interval}$)

Beware Relative Claims.
(a 50% increase in a 0.01% risk is a 0.015% risk)

double check what the error bars represent; think in confidence interval. If it's standard error make it 2x the side in your mind.

Failures of Replication

Not 100% positive an effect doesn't exist, can never truly be. Replication is a flawed term; it's categorical.

However, can be pretty sure about the likely effect size; effects may *exist*, but they're not *meaningful*. We'll cover all of this in the future.



James Vicary and Subliminal advertising. Subliminal things in general.



Original study in *Science*, warm beverages make you feel warm towards people.

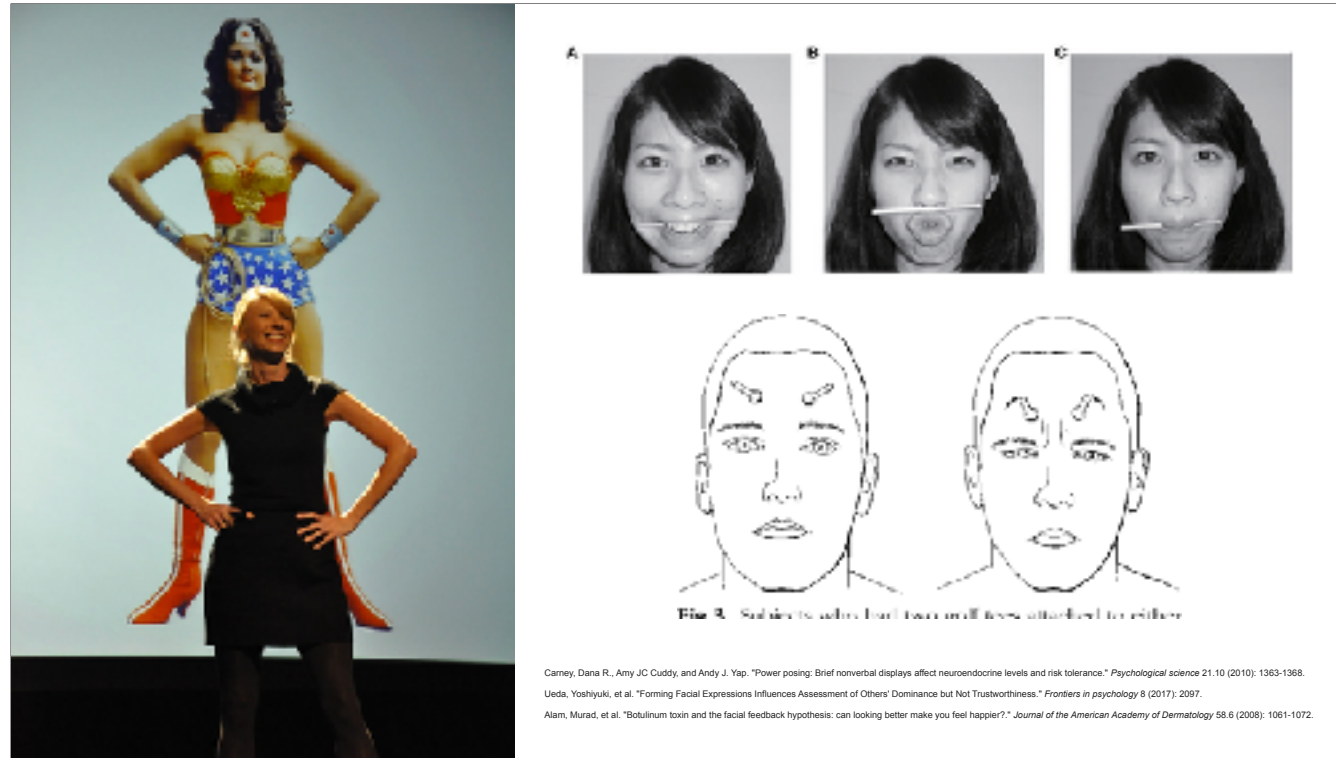
75% of the participants who evaluated a cold pack selected a reward for themselves, whereas 46% of the participants who evaluated a warm pack did the same (analyzed N = 50)

Replications:

S1. N=~300 based off power analysis, pre-registered. S2. N=~300, no effect.

Since 2020, original cited 208 times vs 42 and 35.

Priming in general is dubious— idea that associative links subconsciously drive behavior.



Power pose (good posture makes you more confident)— original first author has come out saying she doesn't believe in it, despite Amy Cuddy still promoting it. Facial feedback— testing whether smile muscle engagement makes you feel happy or not.

Here we're talking about isolated physiological drivers of mood. Telling someone to smile, or reminding yourself to fix your posture, act confidently, and engaging in a ritual that will you believe will make you feel confident, all have an effect— the thing that doesn't have strong support is that there is something intrinsic to the actual physiological change (engaging the muscles) that drives the change in mood.

Larsen, Randy J., Margaret Kasimatis, and Kurt Frey. "Facilitating the furrowed brow: An unobtrusive test of the facial feedback hypothesis applied to unpleasant affect." *Cognition and Emotion* 6.5 (1992): 321-338.

RCTs to Scale: Comprehensive Evidence from Two Nudge Units*

Stefano DellaVigna Elizabeth Linos
UC Berkeley and NBER UC Berkeley

July 2020

Abstract

Nudge interventions - behaviorally-motivated design changes with no financial incentives - have quickly expanded from academic studies to larger implementation in so-called Nudge Units in governments. This provides an opportunity to compare interventions in research studies, versus at scale. We assemble a unique data set of 120 RCTs covering over 23 million individuals, including all trials run by two of the largest Nudge Units in the United States. We compare these trials to a separate sample of nudge trials published in academic journals from two recent meta-analyses. In papers published in academic journals, the average impact of a nudge is very large - an 8.7 percentage point take-up effect, a 33.5% increase over the average control. In the Nudge Unit trials, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.1% increase. We consider five potential channels for this gap: statistical power, selective publication, academic involvement, differences in trial features and in nudge features. Publication bias in the academic journals, exacerbated by

largest RCT of 'nudges' and behavioral economics from US government intervention. 1.4% points; experts thought it'd be 8.1%.

1.4% is meaningful at the scale of countries, and could be worth the effort. 1.4% is unlikely to show up in your n=20 study, and is unlikely to be 'meaningful' at the individual level for intervention design.

“Primeworld”

‘primeworld’ coined by Jesse Singal. Good description.

Contextualized in the history of psychology— Mischel (marshmallow test) wrote 1968’s *Personality and Assessment* claimed that behavior is too cross-situationally inconsistent to be classified with personality traits. Hugely influential. Social psychology and situationism became dominant; trait theory and personality psychology became rare disciplines. Not an accurate worldview.

**Implications of Debunking the “Critical Positivity Ratio” for Humanistic Psychology:
Introduction to Special Issue**

Harris L. Friedman, Nicholas J. L. Brown

First Published March 29, 2018 | Research Article | [Find in PubMed](#) | [Check for updates](#)
<https://doi.org/10.1177/0022167818762227>

[Article information](#) ▾

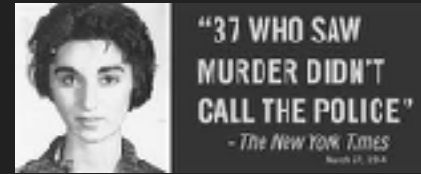


Abstract

An extraordinary claim was made by one of the leading researchers within positive psychology, namely, there is a universal-invariant ratio between positive to negative emotions that serves as a unique tipping point between flourishing and languishing in individuals, marriages, organizations, and other human systems across all cultures and times. Known as the “critical positivity ratio,” this finding was supposedly derived from the famous Lorenz equation in physics by using the mathematics of nonlinear dynamic systems, and was defined precisely as “2.9013.” This exact number was widely touted as a great discovery by many leaders of

‘positivity ratio’ by Barbara Fredrickson, corrected by Nick Brown. Idea that 2.9013 positive to negative feels is a tipping point for all people (as derived from fluid dynamics equations), above which you positive-feedback and flourish, and below which you negative-feedback and languish.

The anti-hedonic set point. Taught for ~8 years as good empirical work. Very wrong. Lots of ‘McMindfulness’ positive psychology out there at the moment.



Even canonical studies from psychology have been re-examined and reinterpreted. Milgram experiments were accurate but his 'agentic state' hypothesis he promoted is pretty clearly inaccurate (that people readily cede their agency and moral accounting to authority); Kitty Genovese case which kicked off 'the bystander effect' was a fraud. Lot's wrong with the stories that have shaped our understanding of human psychology.



Daryl Bem again. Bem is great— he's been very transparent with his data and open about his work. He is a true believer in psychic phenomena.

Paper still not retracted, despite public efforts as recently as this year. In fact, 2015 paper Bem provided a meta-analysis using the latest Bayesian techniques to prove that premonitions are, in fact, still real.

Methodical problems continue... or, perhaps, we're all a little psychic. ;)

Most social psychology research doesn't replicate; revise your priors down and be skeptical.

Trust your intuition and ability to judge the quality of research.

Look at effect sizes (absolute and normalized), not just p-values.

Don't say 'statistically significant' or ** your research.

Report actual p-values, not 'p<alpha'.

Interpret p-values based on your priors; don't confuse PPV (what's the likelihood the effect is real?) with p (how likely to see this data with no effect?)

Assume reported effect sizes are 1/2 to 1/100 what is reported; beware the winner's curse.

Beware of in-vogue research topics; they're more likely to suffer from publication bias and file drawer effects.

Check meta-analyses of journals and authors to inform your priors about publication bias and p-hacking.

Publish all your findings.

Don't HARK (come up with hypotheses after the experiment is over) unless explicitly being *exploratory*. Be explicit about what is a priori hypothesis and what is exploratory.

Don't run multiple small variations of your analyses (beware weird mediating variables).

Don't check your results multiple times after subsets of participants (beware unusual Ns, unusual combinations of subsets of data)

Don't test many hypotheses that could confirm your narrative, and only report the 'winners' (beware narrative/motivated research).

Perform an a priori power analysis.

Pre-register your studies.

Have an explicit, a priori hypothesis and data analysis method; if you don't, be sure to be very explicit that your study is exploratory. You will need to perform a confirmatory study afterwards.

Beware 'N=20' research; guess the power of the study yourself before looking at results and don't trust underpowered studies.

Read the studies you cite.

Check for retractions; check blogs and meta-analyses on your topic of interest.

Look for misleading statistics in the papers; mind the axes, check the error bars, beware relative claims.