Statistical Foundations: Replication (and the failures)

18 October 2021

Modern Research Methods

Midterm

- Available tomorrow at noon on course website
- Due Thursday at noon (via Canvas)

PSYCHOLOGY

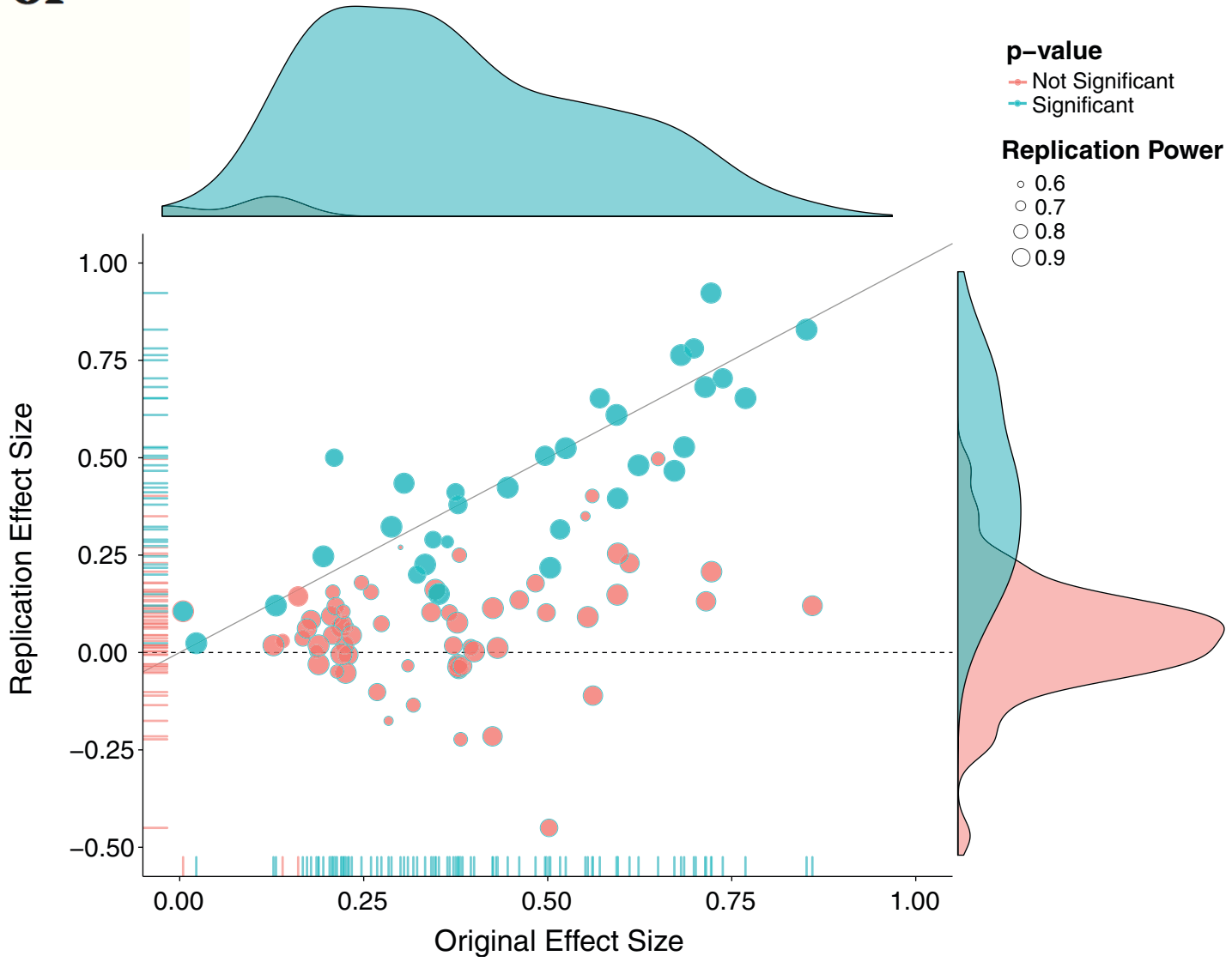
Estimating the reproducibility of psychological science

Open Science Collaboration* (2015)

Conducted replications of 100 psychology effects

Replication effect size half the size of the original, on average.

97 of original studies had $p < .05$; only 37 of replications (47 in CI of original)



Potential reasons for replication failure

1. Fraud
2. Actual change in population effect
3. Error in reporting/analysis
4. Hidden moderator
5. Inadequate materials/description
6. Data-depending analysis (“p-hacking”/“HARKing”)
7. File drawer problem (“publication bias”)
8. Low study precision

Reason #3: Error in reporting/analysis

Current Biology
Retraction

Retraction Notice to: The Negative Association between Religiousness and Children's Altruism across the World

Jean Decety,* Jason M. Cowell, Kang Lee, Randa Mahasneh, Susan Malcolm-Smith, Bilge Selcuk, and Xinyue Zhou
*Correspondence: decety@uchicago.edu
https://doi.org/10.1016/j.cub.2019.07.030

(Current Biology *25*, 2951–2955; November 16, 2015)

In our paper, we reported cross-cultural differences in how the religious environment of a child negatively impacted their sharing, their judgments of the actions of others, and how their parents evaluated them. An error in this article, our incorrect inclusion of country of origin as a covariate in many analyses, was pointed out in a correspondence from Shariff, Willard, Muthukrishna, Kramer, and Henrich (https://doi.org/10.1016/j.cub.2016.06.031). When we reanalyzed these data to correct this error, we found that country of origin, rather than religious affiliation, is the primary predictor of several of the outcomes. While our title finding that increased household religiousness predicts less sharing in children remains significant, we feel it necessary to explicitly correct the scientific record, and we are therefore retracting the article. We apologize to the scientific community for any inconvenience caused.

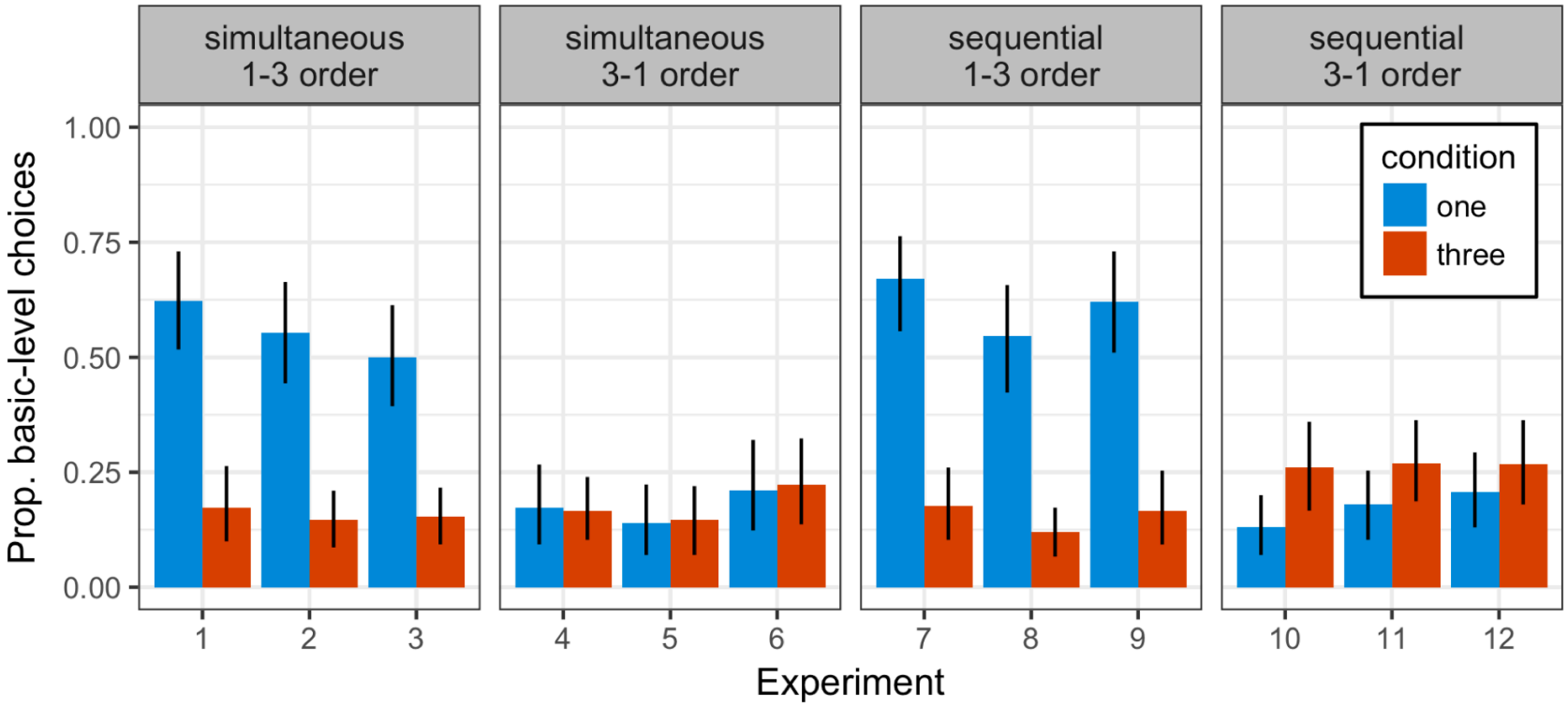
Coded country as continuous measure, rather than categorical.

1 – US
2 – Canada
3 – Japan

Reason #4: Hidden moderator

Replication of
Xu and Tenenbaum (2007)

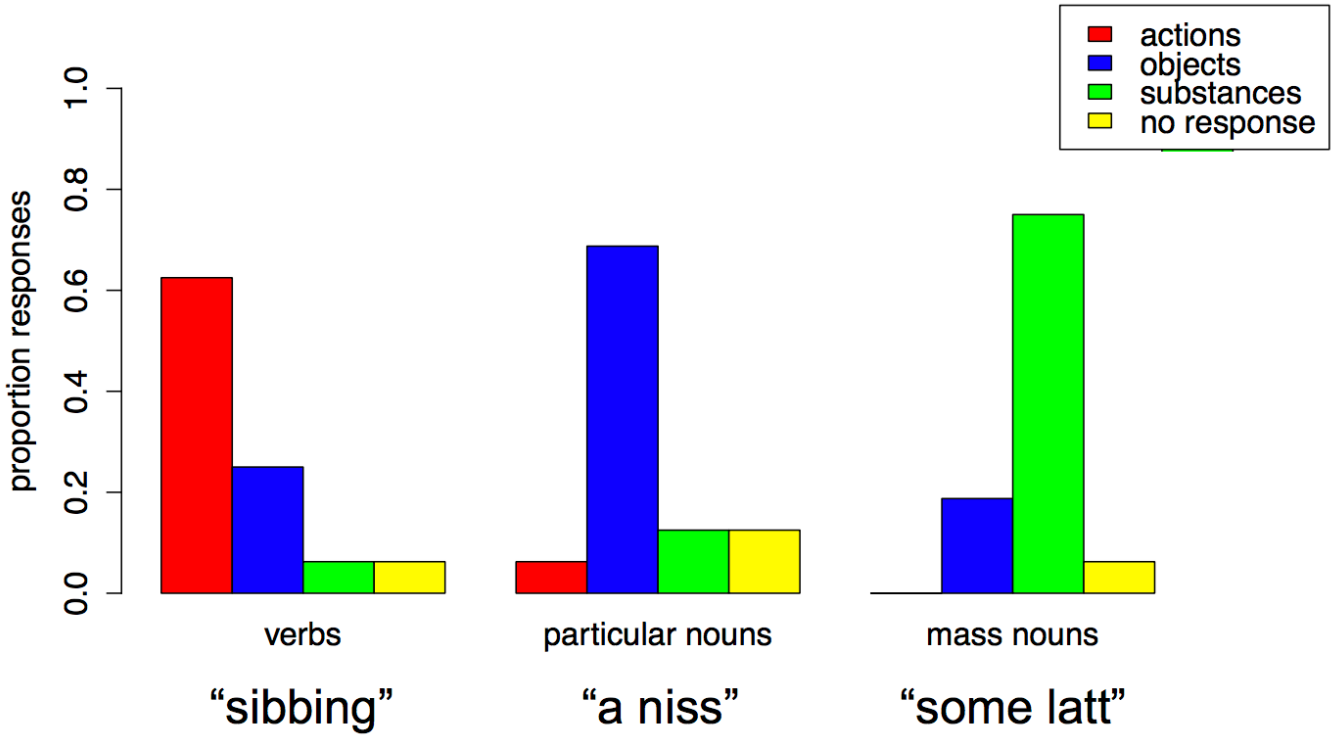
Replication of
Spencer et al. (2011)



(Lewis & Frank, 2016)

Reason #5: Inadequate materials/description

Syntactic Bootstrapping (Brown, 1957)



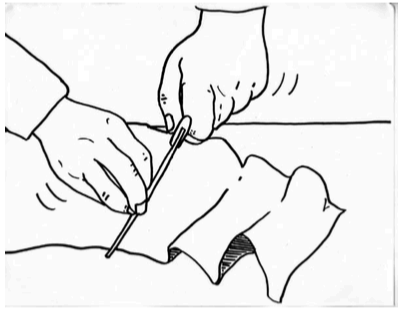
Reason #5: Inadequate materials/description

METHOD

The experiment involved three sets of four pictures each.[2] One of these sets will be described in detail. The first picture in the set shows a pair of hands performing a kneading sort of motion, with a mass of red confetti-like material which is piled into and overflowing a blue-and-white striped container that is round and low in shape. The important features of the picture are the kneading action, the red mass, and the blue-and-white round container. The motion would ordinarily be named with a verb (like "kneading"), the mass with a mass noun (like "confetti"), and the container with a particular noun (like "container"). It was assumed that children would have no readily available names for any of these conceptions. Each of the remaining three pictures of this set exactly reproduced one of the three salient features of the first picture, either the motion, the mass, or the container. In order to represent the motion a second time it was necessary to show also a mass and a container. However, the mass was here painted yellow so as not to duplicate the original, and the container was of a different size, shape, and color from the original. The other two sets of pictures involved different content, but always an action, a mass

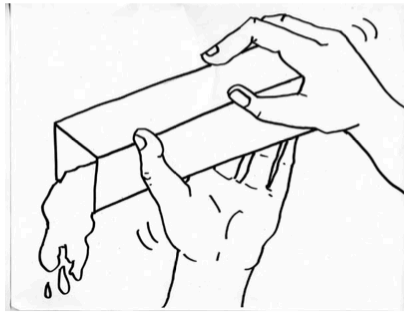
Do you know what it means to sib?

In this picture, you can see sibbing.



Do you know what a niss is?

In this picture, you can see a niss.



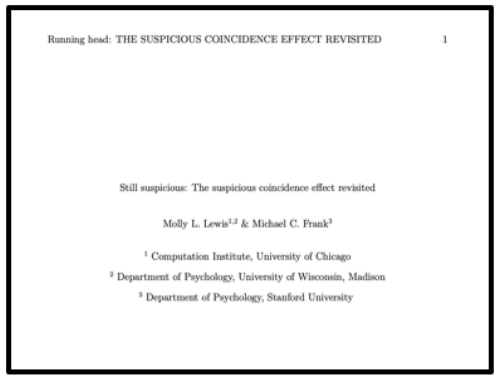
Have you ever seen any latt?

In this picture, you can see some latt.

Reason #6: Data-dependent analysis

Historically, need $p<.05$ for paper to get published

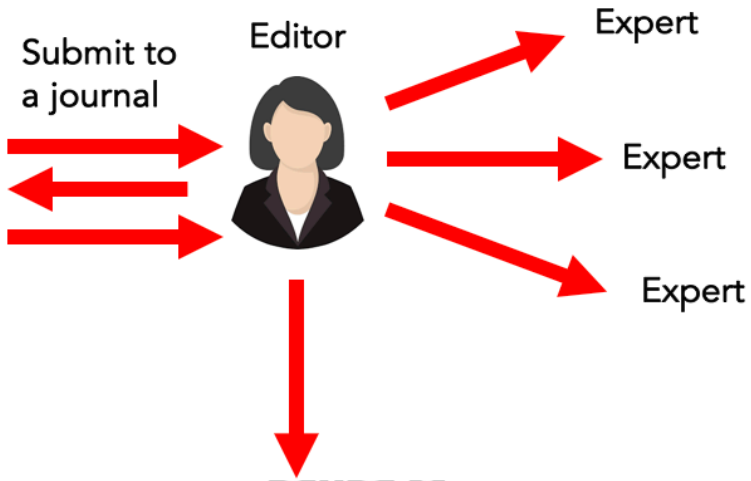
Peer Review Process


Running head: THE SUSPICIOUS COINCIDENCE EFFECT REVISITED 1

Still suspicious: The suspicious coincidence effect revisited

Molly L. Lewis[1,2] & Michael C. Frank[3]

[1] Computation Institute, University of Chicago
[2] Department of Psychology, University of Wisconsin, Madison
[3] Department of Psychology, Stanford University



Preregistered Direct Replication

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Still Suspicious: The Suspicious-Coincidence Effect Revisited

Psychological Science
2018, Vol. 29(12) 2039–2047
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797618794931
www.psychologicalscience.org/PS
SAGE

Molly L. Lewis[1,2] and Michael C. Frank[3]
[1]Computation Institute, University of Chicago; [2]Department of Psychology, University of Wisconsin–Madison; and [3]Department of Psychology, Stanford University

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Choosing your analysis based on seeing your data/the outcome of a test ("analytic flexibility")

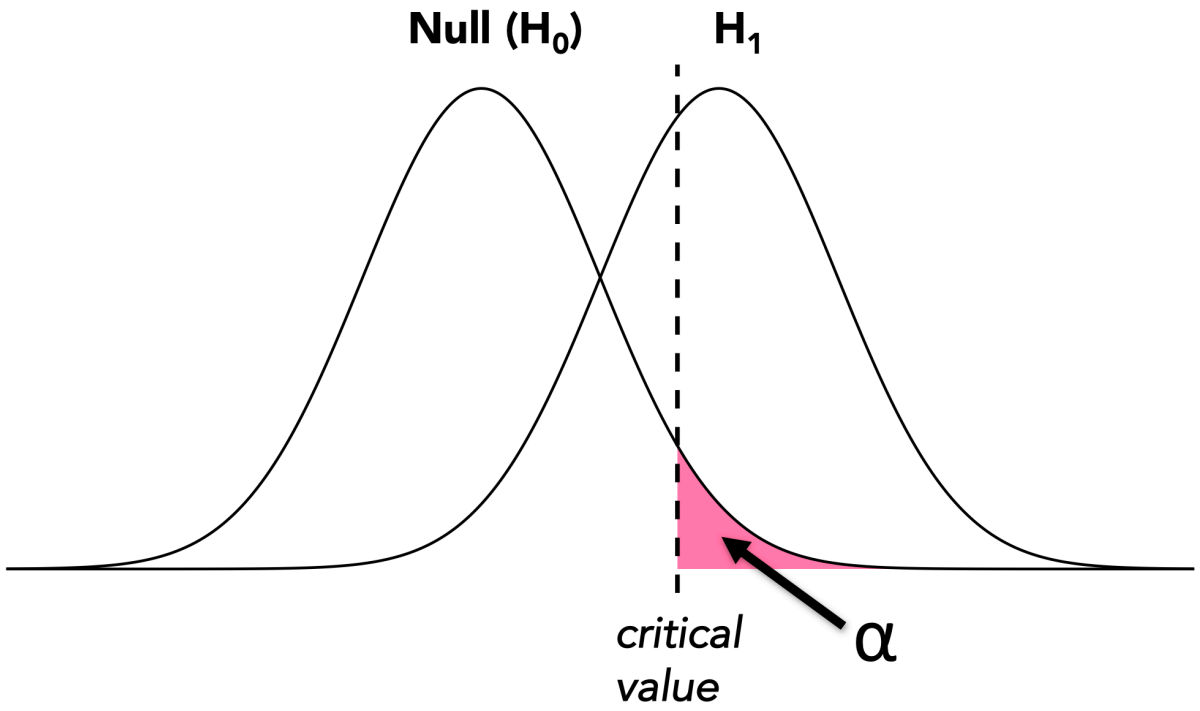
"*p*-hacking"/"Questionable research practices" (QRP)

"…it is unacceptably easy to publish *statistically significant* evidence consistent with *any* hypothesis" -- Simmons, Nelson, & Simonsohn, 2011

"Researcher degrees of freedom" -- flexibility in data collection, analysis, and reporting

Why is *p*-hacking bad?

p-values aren't veridical read outs of truth, they're probabilities, and therefore depend on the number of "samples" you take.



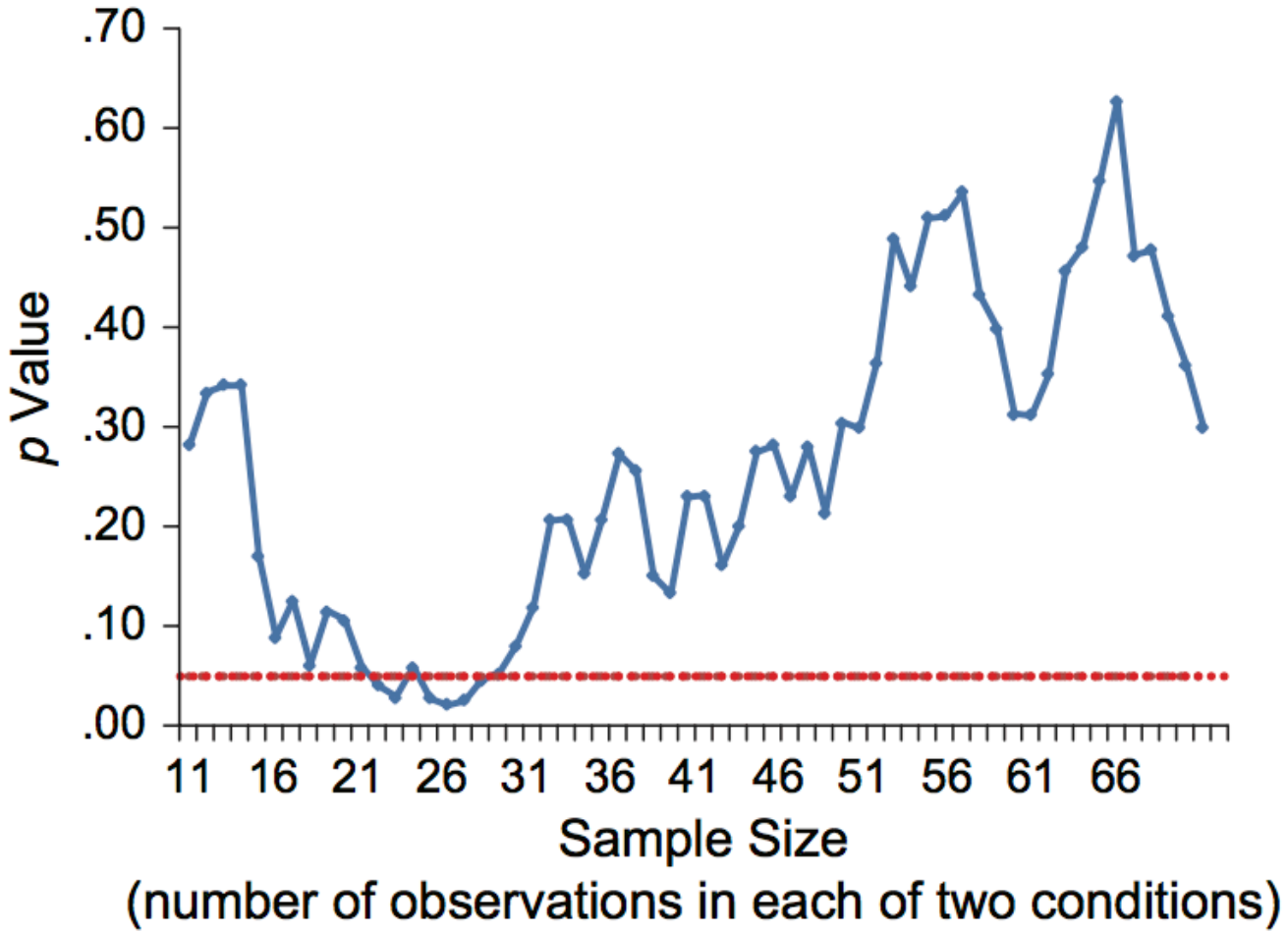


Fig. 2. Illustrative simulation of *p* values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a *t* test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

(Simmons, et al. 2011)

Examples of p-hacking

Collect more data?

Should some observations be excluded? Which ones?

Which conditions should be combined with which ones?

Which measures should we analyze? Should we transform the measure?

Which control variables should we consider?

Should we include pilot data?

Original report

RESEARCH | *REPORTS*

EDUCATION

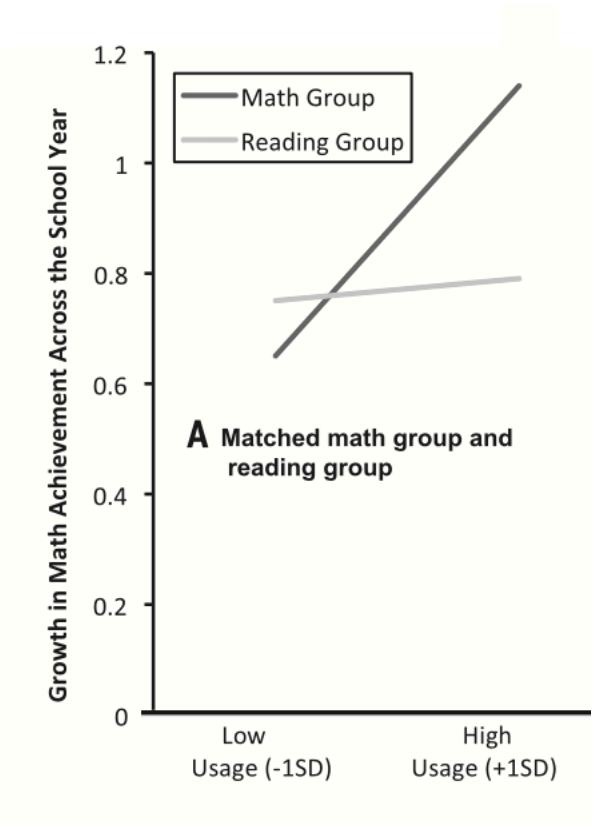
Math at home adds up to achievement in school

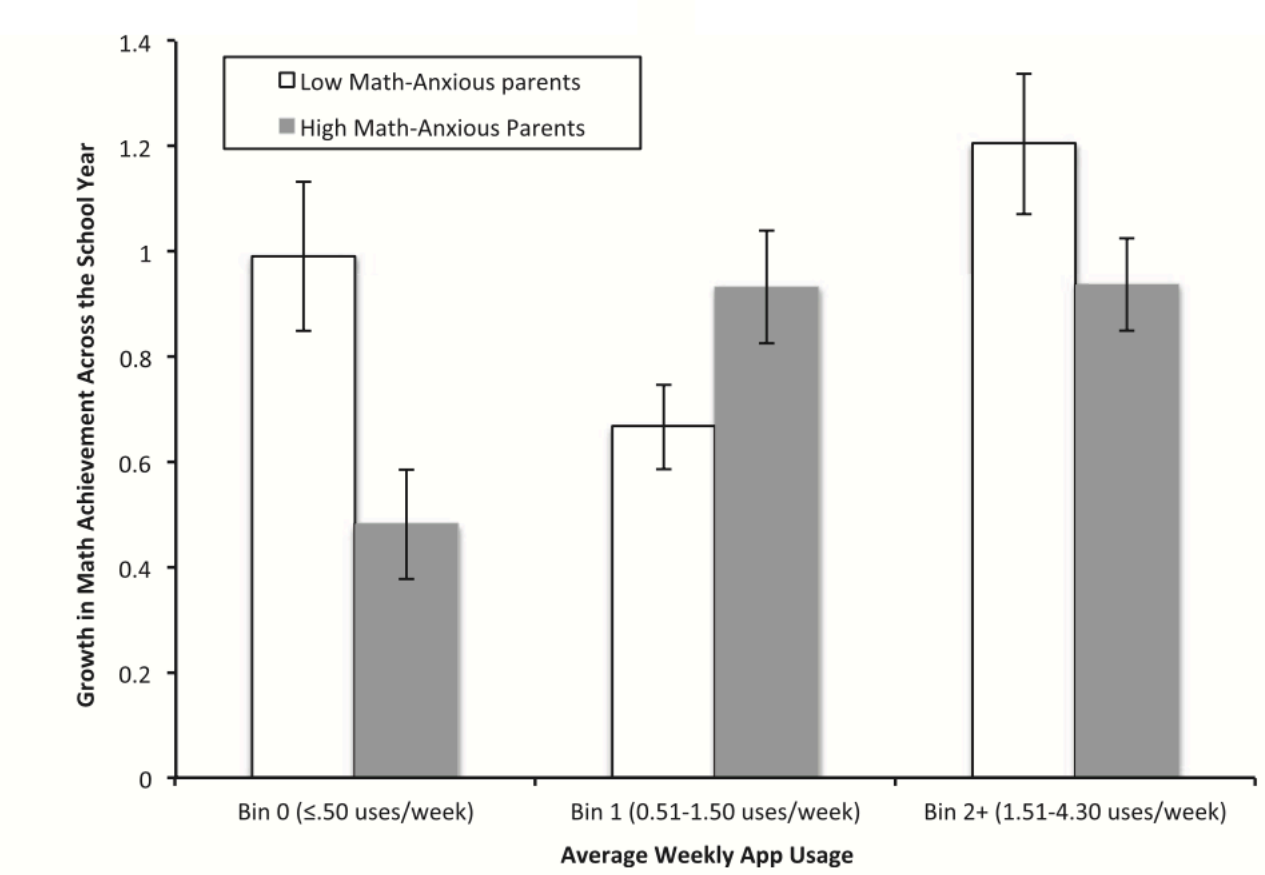
Talia Berkowitz,* Marjorie W. Schaeffer,* Erin A. Maloney, Lori Peterson, Courtney Gregor, Susan C. Levine,† Sian L. Beilock†

With a randomized field experiment of 587 first-graders, we tested an educational intervention designed to promote interactions between children and parents relating to math. We predicted that increasing math activities at home would increase children's math achievement at school. We tested this prediction by having children engage in math story time with their parents. The intervention, short numerical story problems delivered through an iPad app, significantly increased children's math achievement across the school year compared to a reading (control) group, especially for children whose parents are habitually anxious about math. Brief, high-quality parent-child interactions about math at home help break the intergenerational cycle of low math achievement.

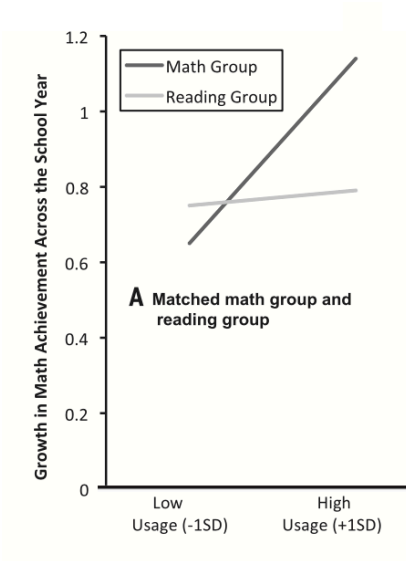
What analysis would you do to test their hypothesis?

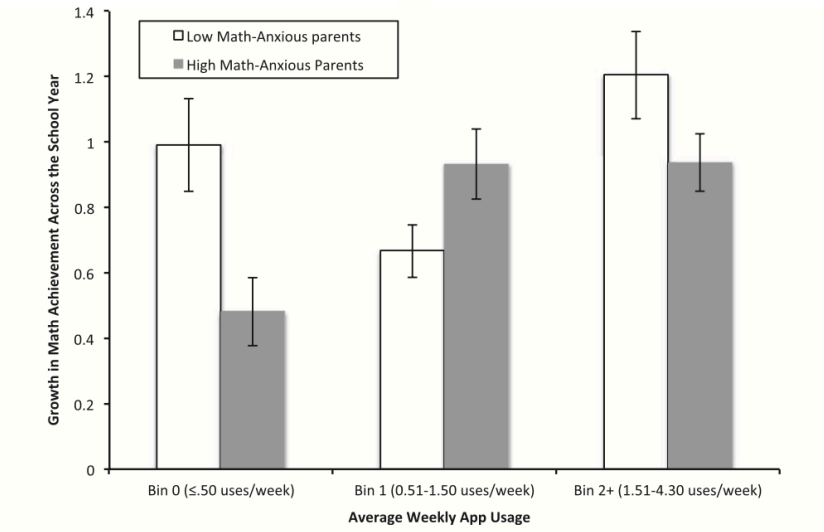
Data

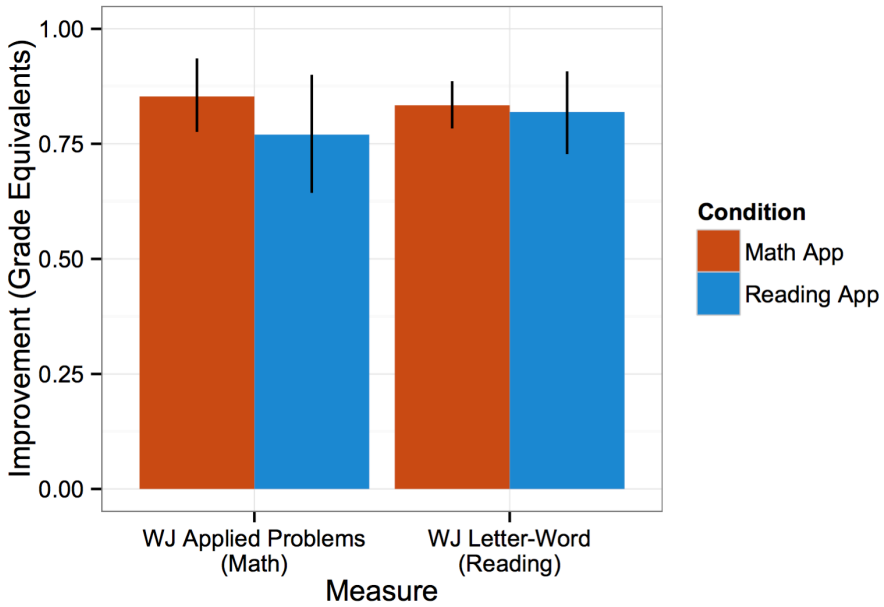

Growth in Math Achievement Across the School Year
1.2
1
0.8
0.6
0.4
0.2
0
Math Group
Reading Group
A Matched math group and reading group
Low Usage (-1SD)
High Usage (+1SD)

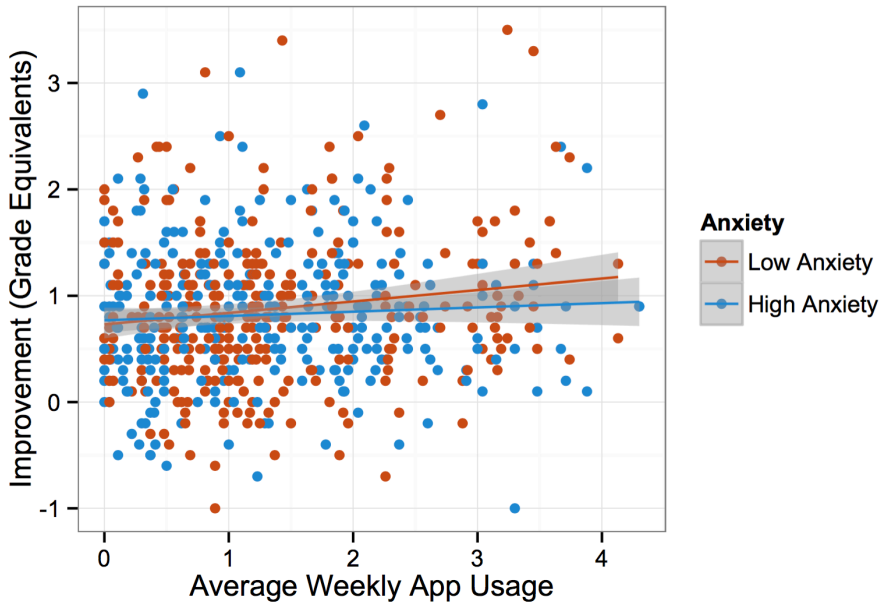

Growth in Math Achievement Across the School Year
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Low Math-Anxious parents
High Math-Anxious Parents
Bin 0 (≤.50 uses/week)
Bin 1 (0.51-1.50 uses/week)
Bin 2+ (1.51-4.30 uses/week)
Average Weekly App Usage

Re-Analysis



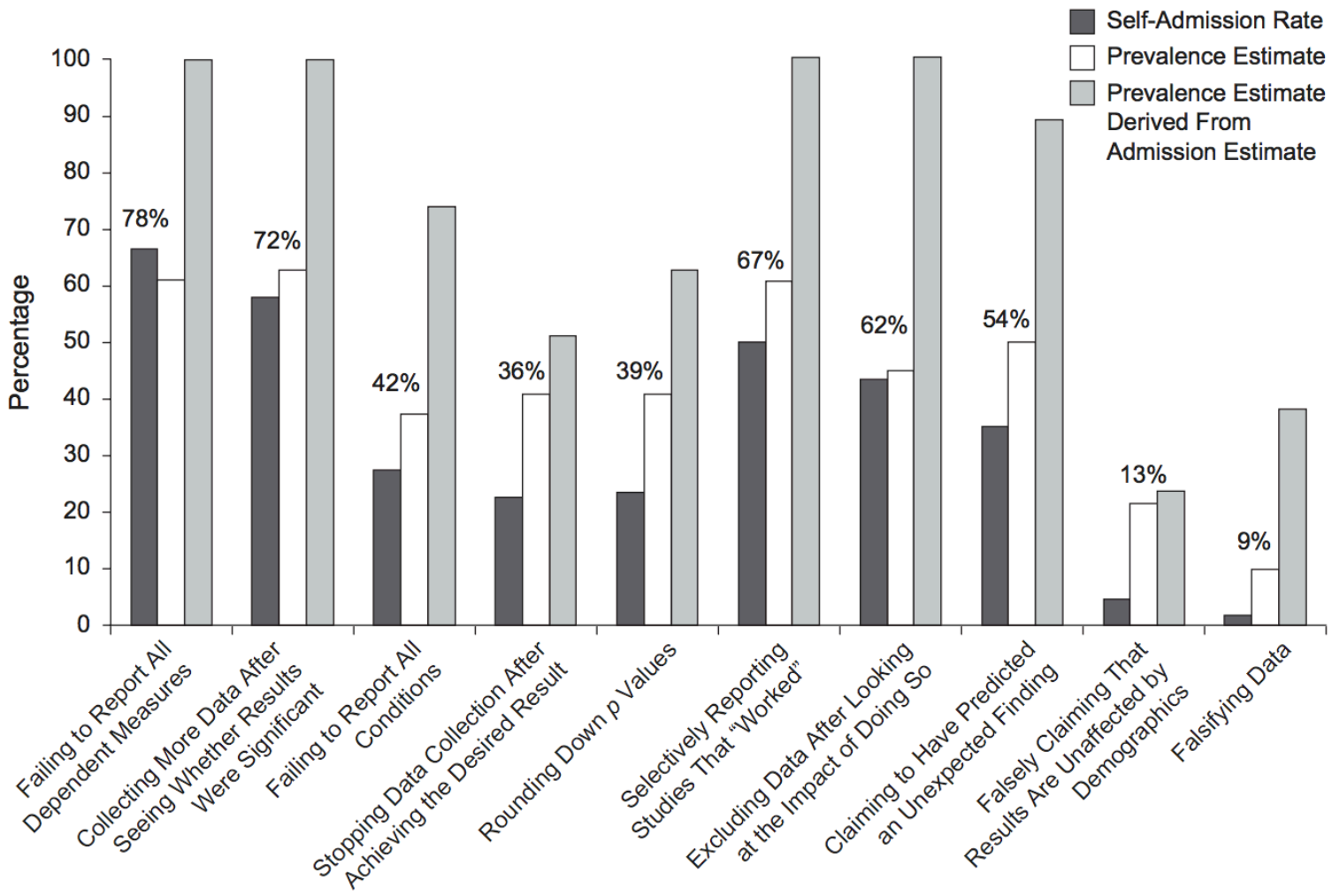






(Frank, 2016)

How common are questionable research practices?



Very common.

(John, Lowenstein, & Prelec, 2012)

Reason #7: File Drawer Problem (or, "publication bias")

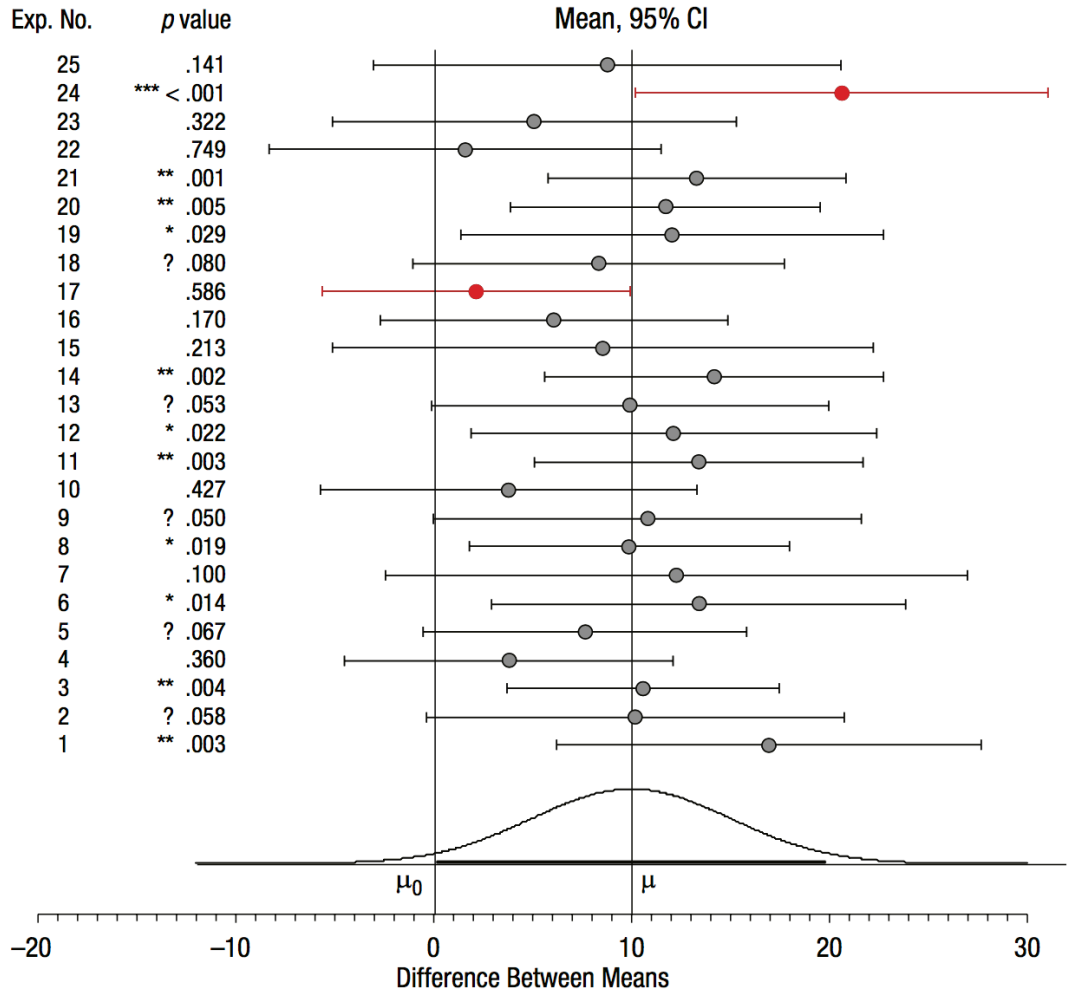
Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

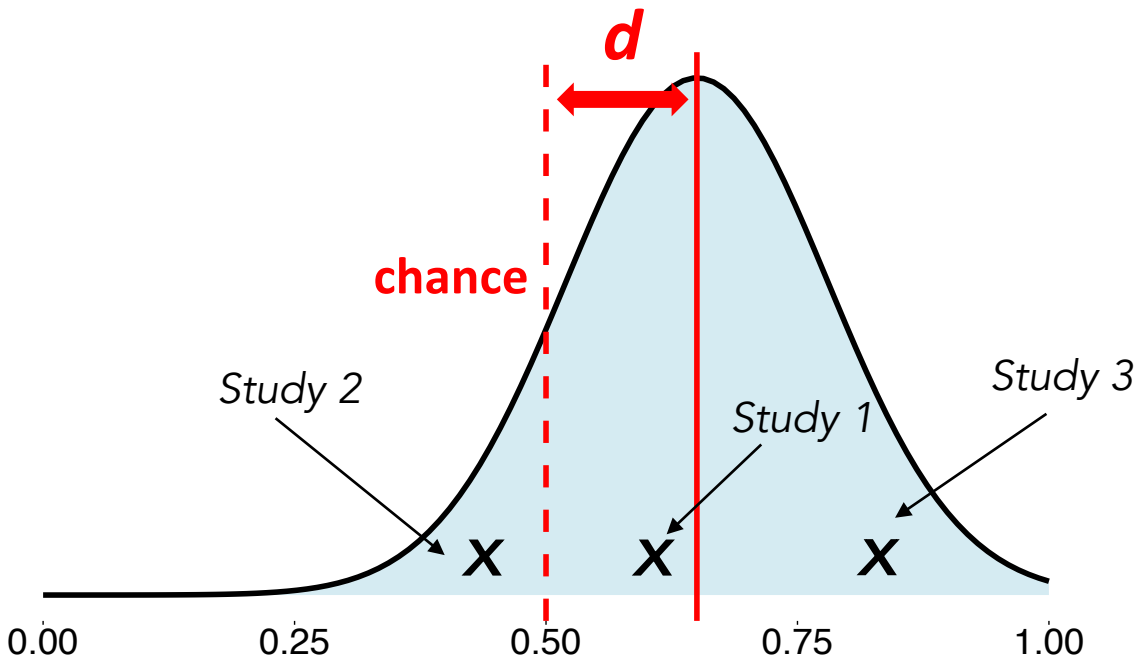
Historically, researchers have only been able to publish results if $p < .05$.

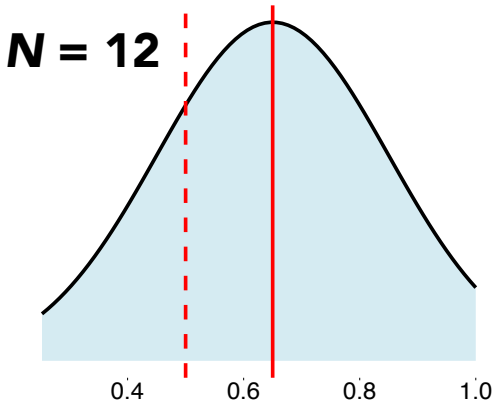
But, they run many studies, and put the studies that "fail" in their file drawer.

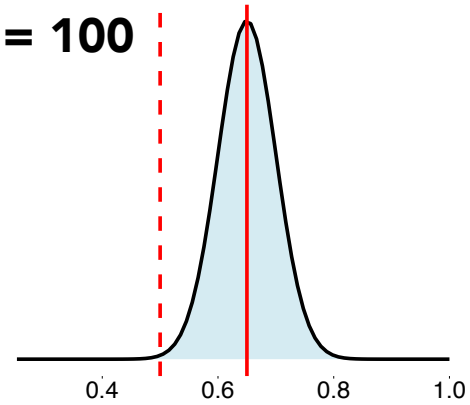


(Cumming, 2014)

Reason 8: Low study precision







Some expected variability in effect size due to sample size; less variability with larger sample sizes.

Potential reasons for replication failure

1. Fraud
2. Actual change in population effect
3. Error in reporting/analysis
4. Hidden moderator
5. Inadequate materials/description
6. Data-depending analysis (“p-hacking”/“HARKing”)
7. File drawer problem (“publication bias”)
8. Low study precision

Next time:

- Replication solutions