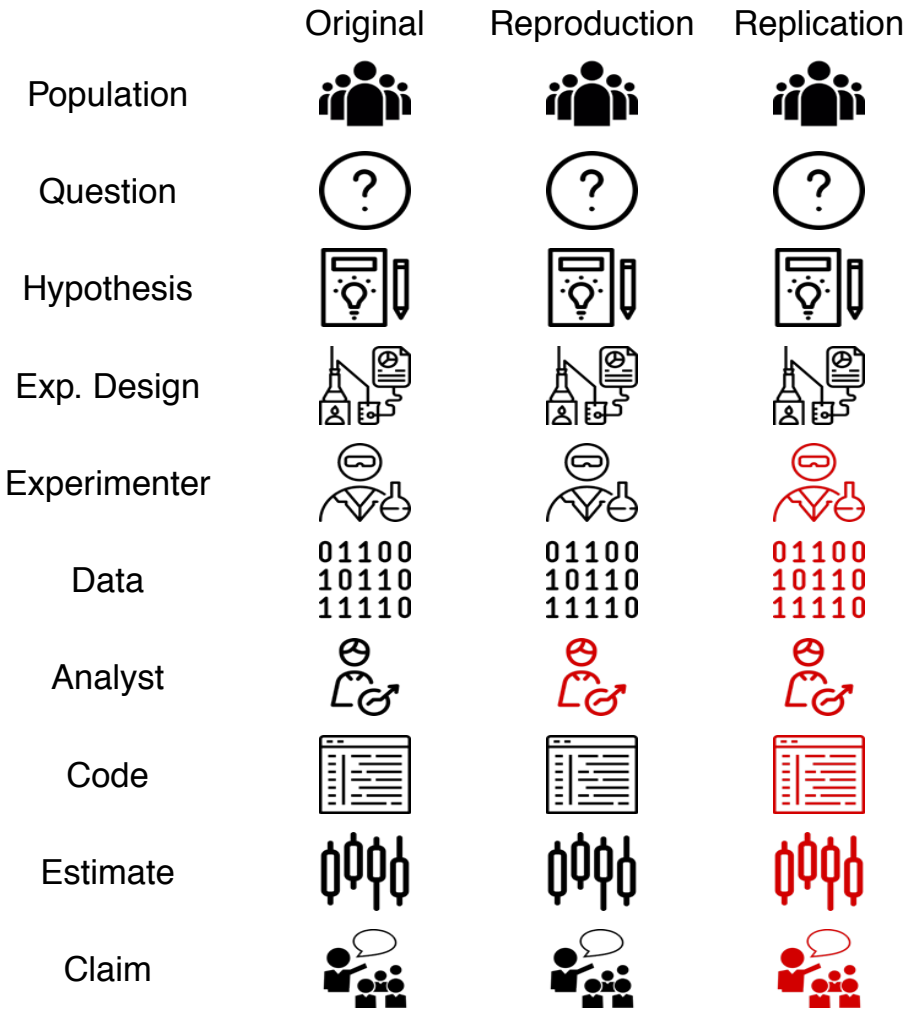
Statistical Foundations: Replication Solutions

4 March 2020

Modern Research Methods

Replications



REPLICATE = Get same result with a new dataset

(Patil, Peng, & Leek, 2019)

An aside – Replications: Who and what

Who does replications?

- As a first step in a project where you're building on an effect in the literature
- Large scale coordinated efforts (e.g., OSF project)
- Students in methods courses (Frank & Saxe, 2012)
- Journals are starting to become more amenable to publishing straight replications.

What counts as a replication?

- No replication is ever exact (different time, experimenter, lighting, etc.)
- In principle, you should think of a close replication as having all the same theoretically relevant methodological choices
- But, ultimately this is subjective and a continuum

"Replication crisis"

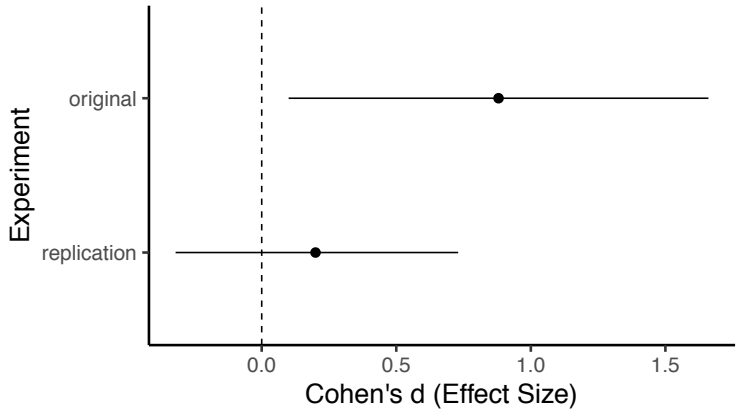
Research Article

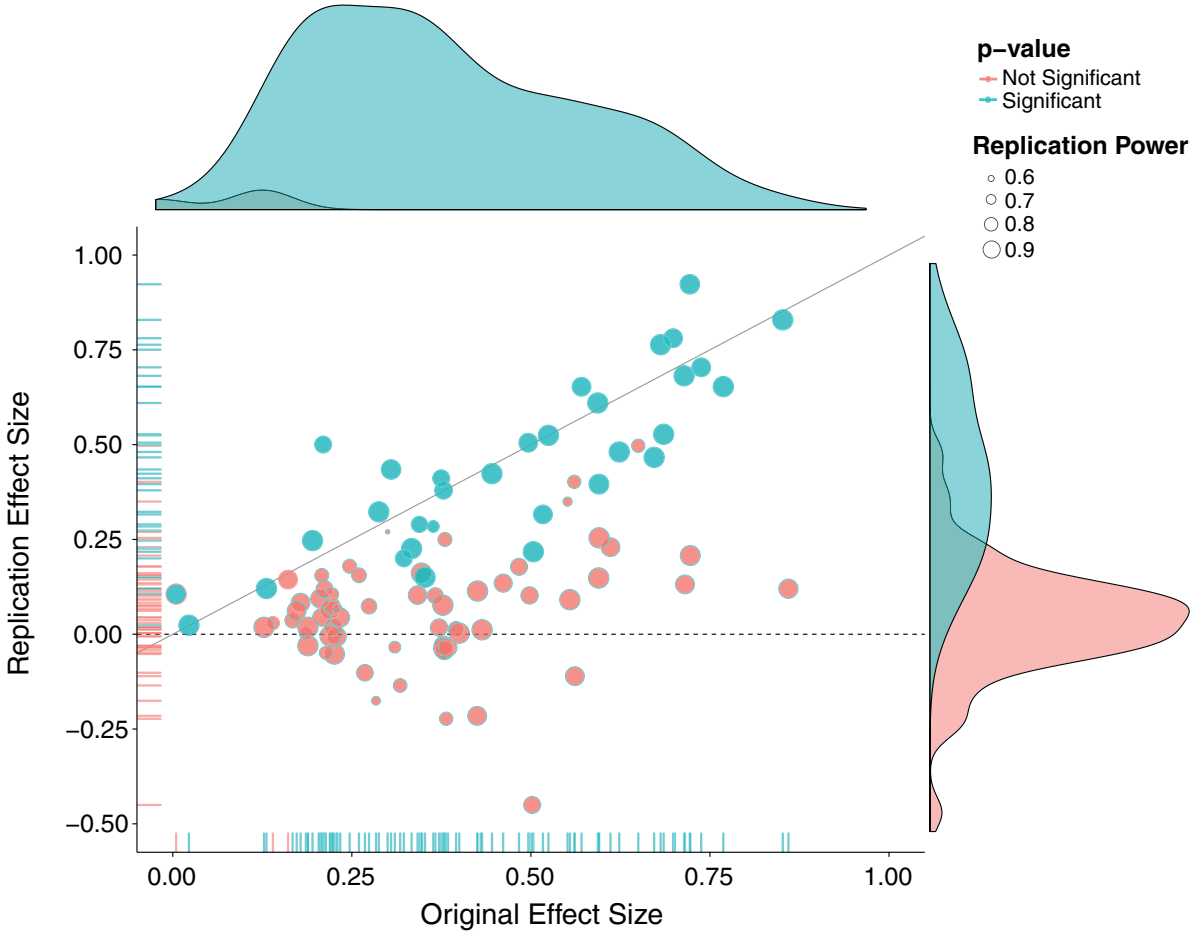
The Value of Believing in Free Will

Encouraging a Belief in Determinism Increases Cheating

Kathleen D. Vohs[1] and Jonathan W. Schooler[2]

[1]Department of Marketing, Carlson School of Management, University of Minnesota, and [2]Department of Psychology, University of British Columbia





Reasons for failed replications

1. Data fraud
2. Analysis/reporting errors
3. Change in population effect size
4. Hidden moderators
5. File drawer
6. Data-dependent analysis (*p*-hacking)

Reason #6: Data-dependent analysis

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

"*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

Collect more data?

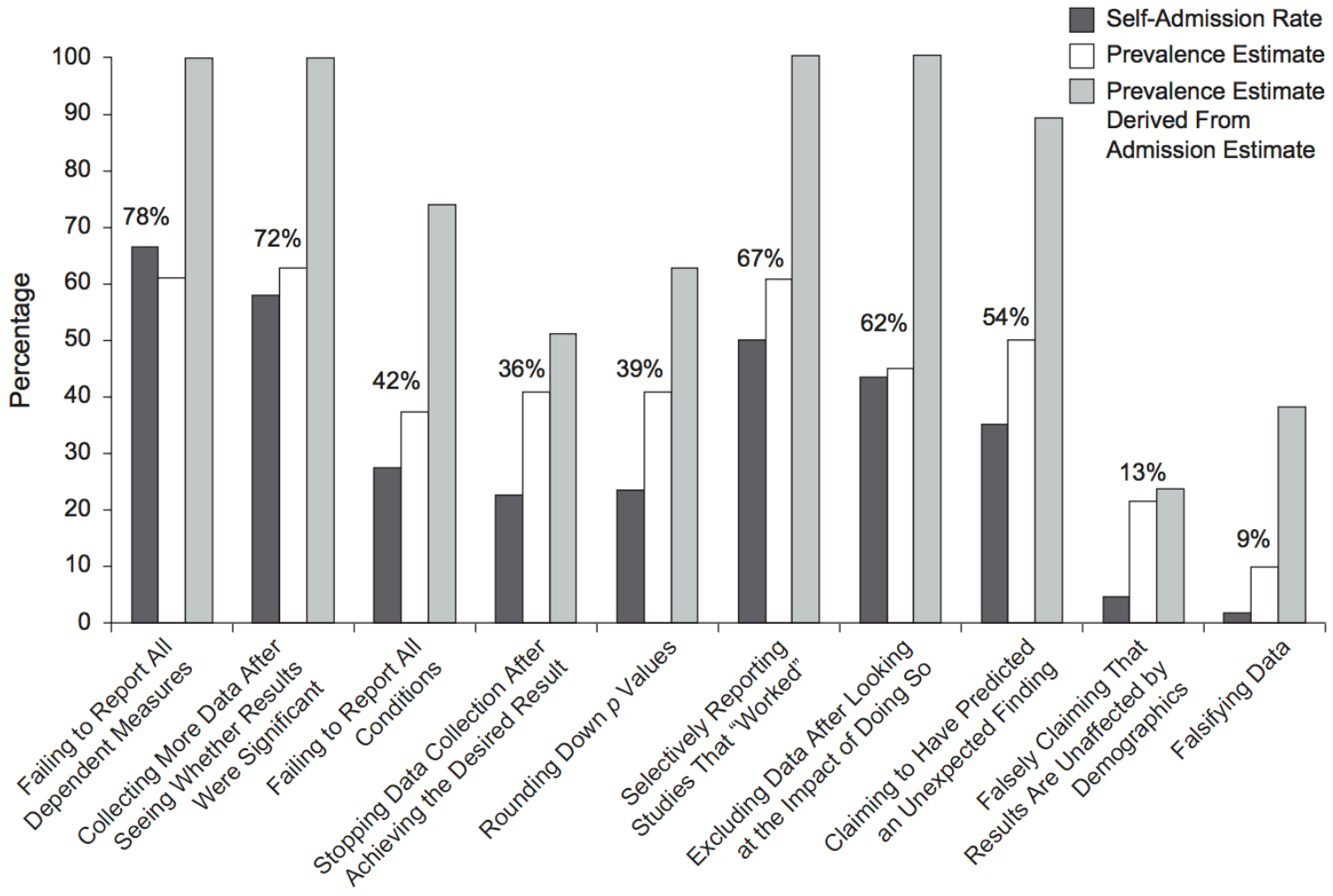
Should some observations be excluded? Which ones?

Which conditions should be combined with which ones?

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

Which control variables should we consider?

How common are QRPs?



Very common.

(John, Lowenstein, & Prelec, 2012)

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.

coin factory

H_0 weight is .5

H_1 weight is not .5

	Retain H_0	**Reject H_0**
H_0 is true	correct	Error (Type I)
H_0 is false	Error (Type II)	correct

Original report

RESEARCH | *REPORTS*

EDUCATION

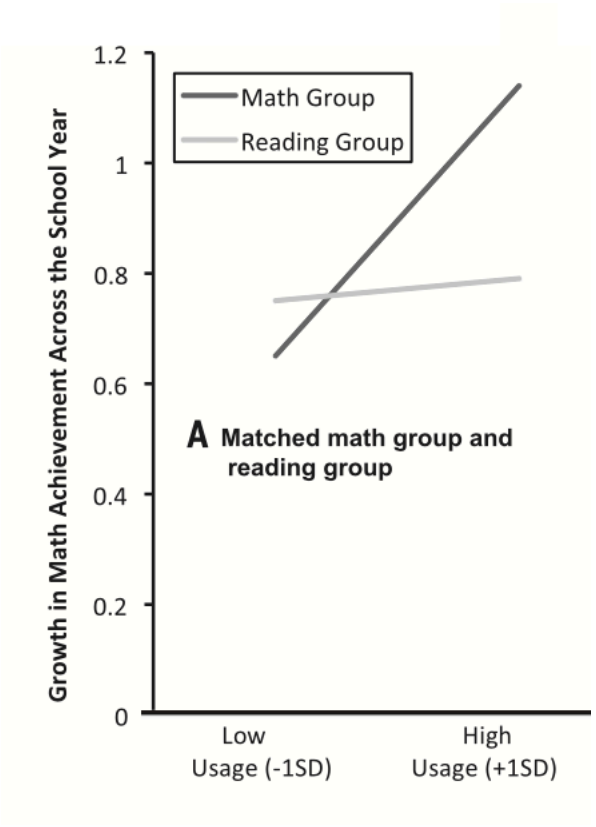
Math at home adds up to achievement in school

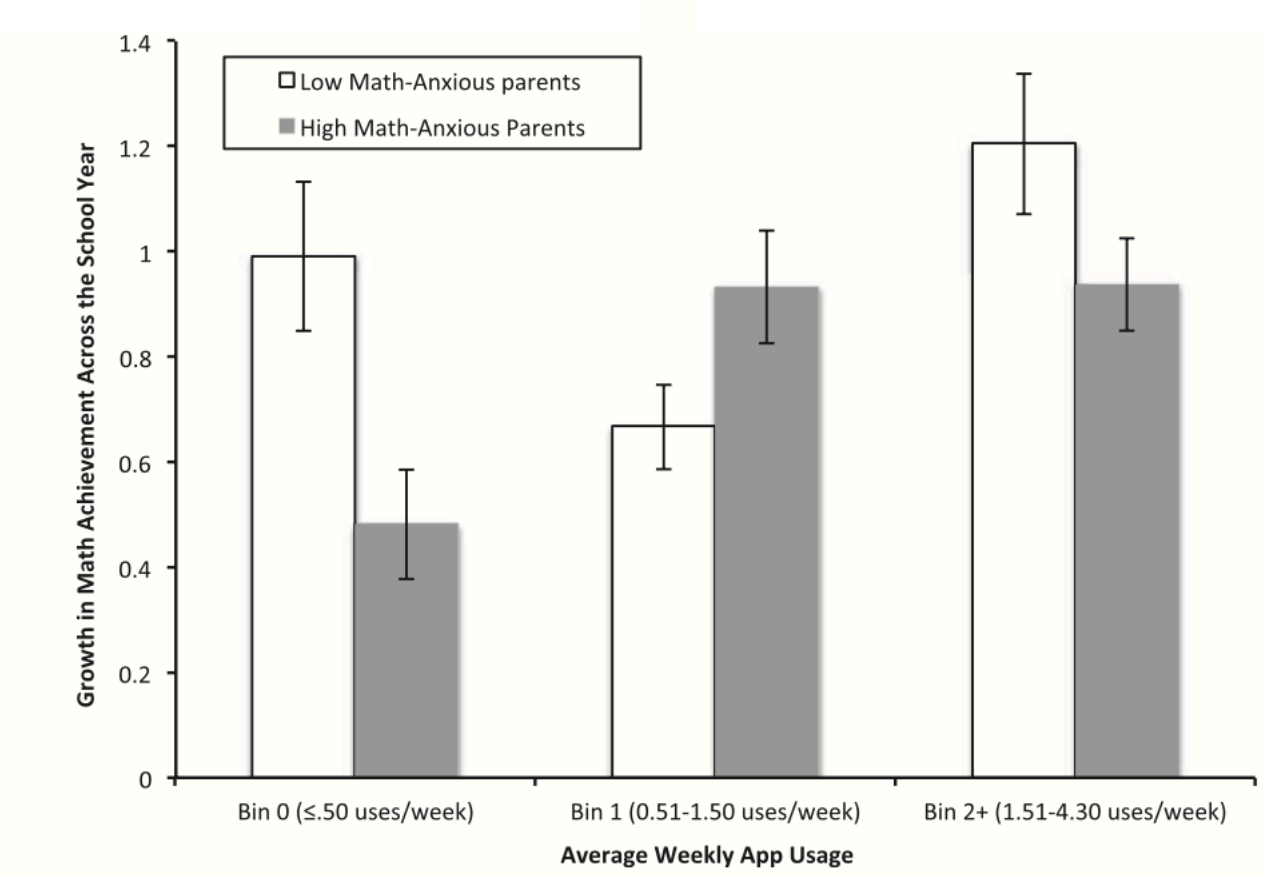
Talia Berkowitz,* Marjorie W. Schaeffer,* Erin A. Maloney, Lori Peterson, Courtney Gregor, Susan C. Levine,† Sian L. Beilock†

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.

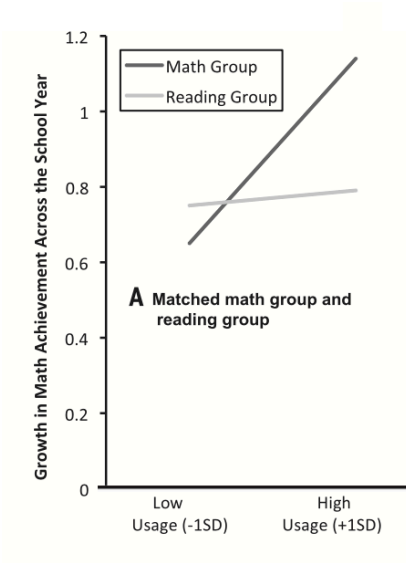
What analysis would you do to test their hypothesis?

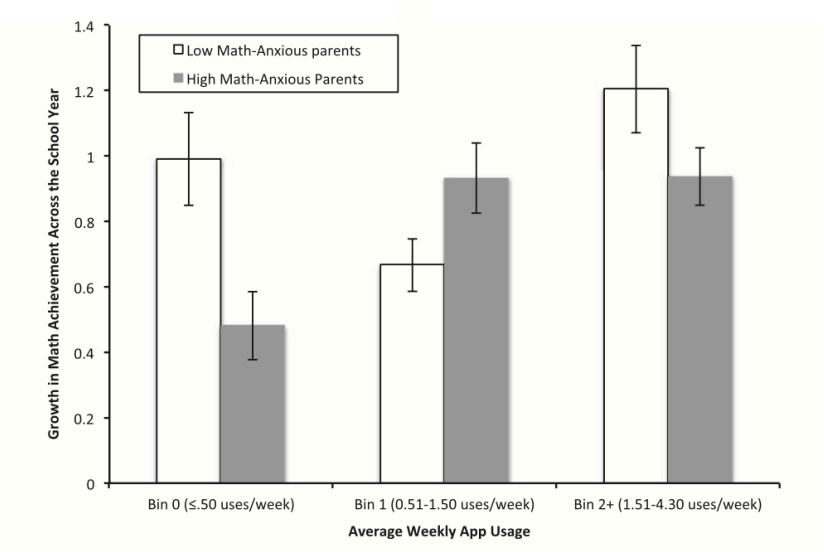
Data

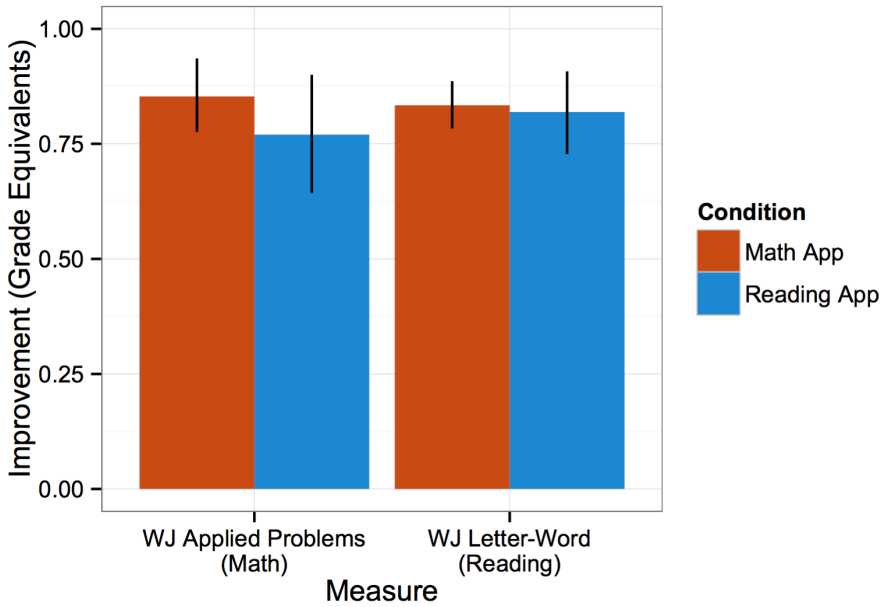

Growth in Math Achievement Across the School Year
1.2
1
0.8
0.6
0.4
0.2
0
Math Group
Reading Group
A Matched math group and reading group
Low Usage (-1SD)
High Usage (+1SD)

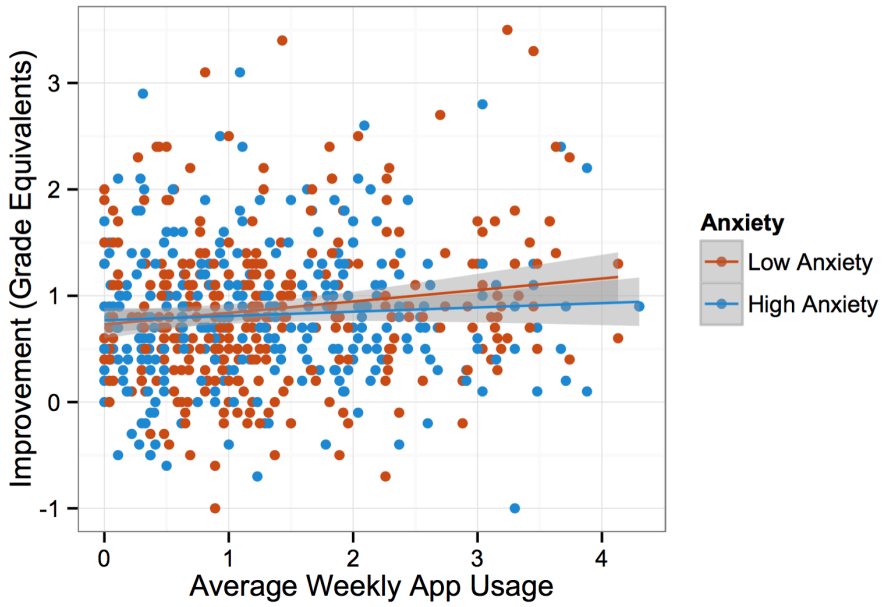

Growth in Math Achievement Across the School Year
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Low Math-Anxious parents
High Math-Anxious Parents
Bin 0 (≤.50 uses/week)
Bin 1 (0.51-1.50 uses/week)
Bin 2+ (1.51-4.30 uses/week)
Average Weekly App Usage

Re-Analysis









(Frank, 2016)

Try your hand at *p*-hacking!

http://shinyapps.org/apps/p-hacker/

Goal: get a publishable finding (i.e. $p < .05$)

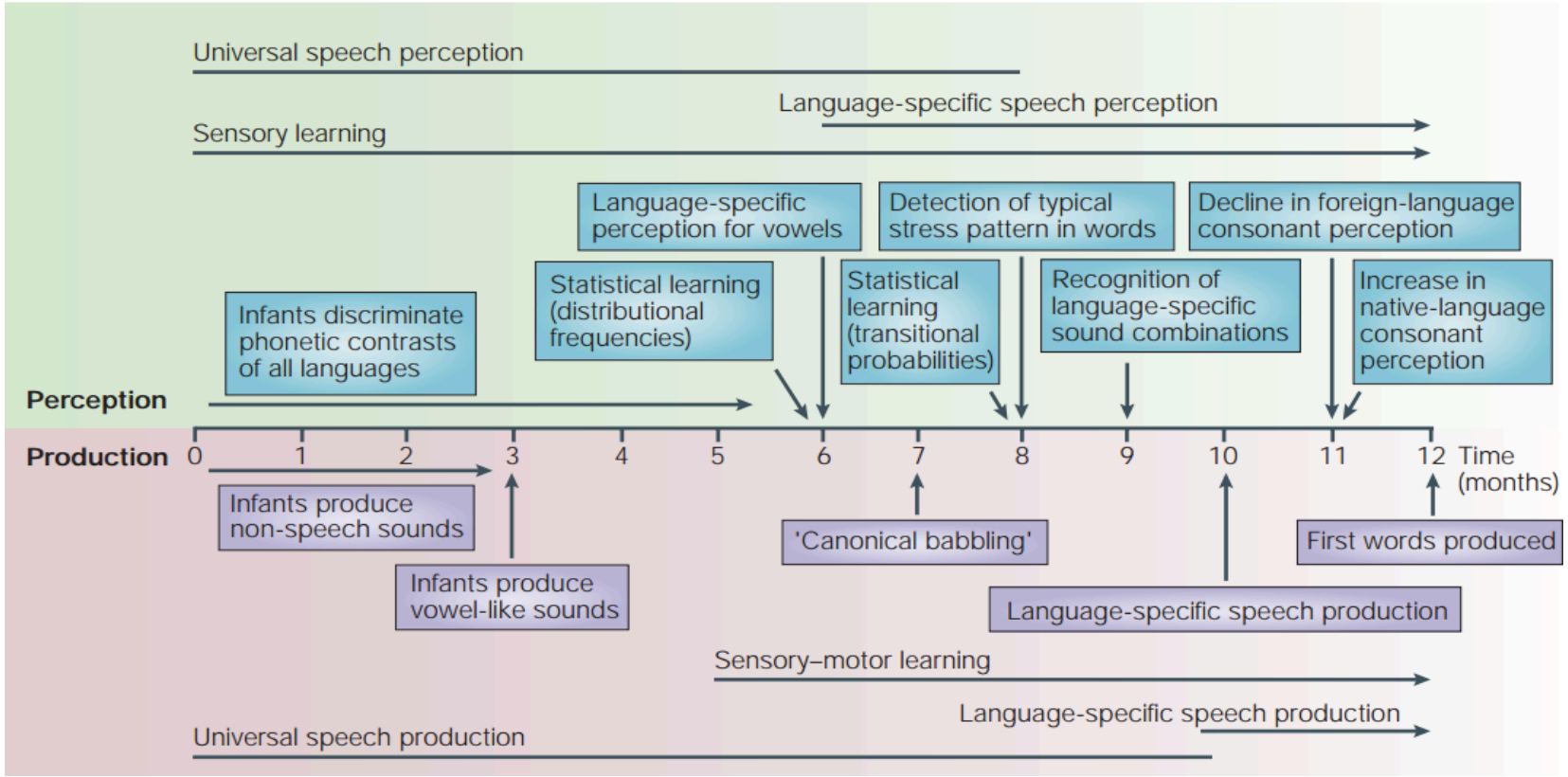
- How many significant p-values can you get?
- What things can you do to increase the likelihood of getting a significant *p*-value?
- What factors make it more likely your *p*-value will be significant?

Reasons for failed replications

1. Data fraud
2. Analysis/reporting errors
3. Change in population effect size
4. Hidden moderators
5. File drawer
6. Data-dependent analysis (*p*-hacking)

Why does replicability matter?

Our goal as scientists is to build predictive theories



But, if the experiments our theories are built on aren't real, then our theories are bad too.

Kuhl (2004)

How can we increase replicability?

Solution

Reproducibility practices

1. Data fraud
2. Analysis/reporting errors
3. Change in population effect size
4. Hidden moderators

Strategies for reducing rates for failed replicates due to reasons that increase Type I error

5. File drawer
6. Data-dependent analysis (*p*-hacking)

Two Solutions

1. Pre-registration – register your hypothesis and analysis plan publically before you collecting your data

2. Registered reports – write the paper and have it reviewed before you collect your data.

https://www.youtube.com/watch?v=LYfZr5poCkQ

Pre-registration

- OSF great tool for registration (essentially, time-stamping)
- All levels of specificity
  - AsPredicted template ("prereg-lite") - https://aspredicted.org/
  - All the way through full analytic script capture

- Example preregistration from Lewis & Frank, 2018: https://osf.io/yekhj/

Preregistration – in sum:

- It costs nothing and makes you feel good.
- If you’re running a study, just try it.
- It'll make you feel like a scientist.

Slide adopted from
Mike Frank

Registered Report

Are the hypotheses well founded?

Are the methods and proposed analyses feasible and sufficiently detailed?

Have the authors included sufficient positive controls to confirm that the study will provide a fair test?

"provisionally accepted"

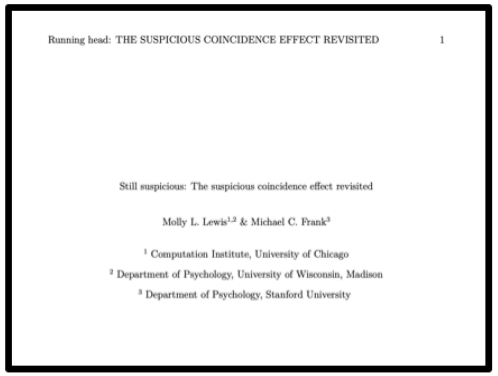
Did the authors follow the approved protocol?

Did positive controls succeed?

Are the conclusions justified by the data?

Slide adopted from Chris Chambers

c.f. Typical peer-review process

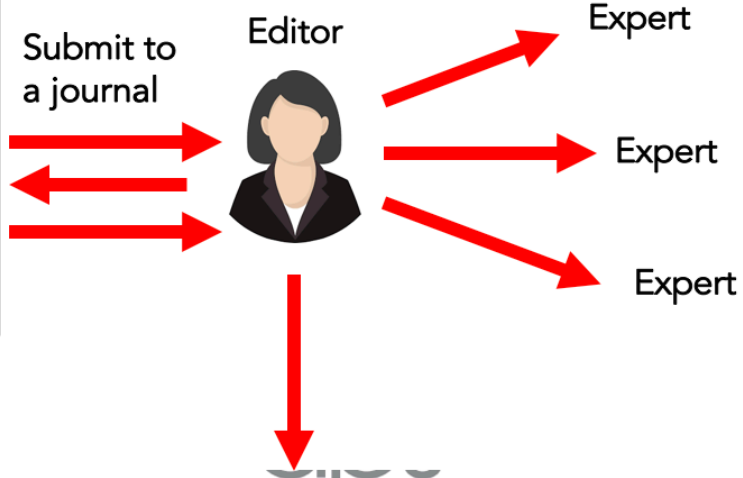


Running head: THE SUSPICIOUS COINCIDENCE EFFECT REVISITED 1

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

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

Two Solutions

1. Pre-registration – register your hypothesis and analysis plan publically before you collecting your data

2. Registered reports – write the paper and have it reviewed before you collect your data.

- Designed to reduce "questionable research practices", like *p*-hacking.
- Thereby, reducing Type I error
- And **increasing replicability**

After Spring Break

- Run a replication of our own study online, using Amazon Mechanical Turk
- Then: Meta-analysis!
- A method for aggregating across effect sizes from individual studies
- “Quantitative literature review”