Statistical Foundations: Replication (and the failures)

13 October 2021

Modern Research Methods

Interpreting confidence intervals

- Interpretation little tricky
- "if we replicated the experiment over and over again and computed a 95% CI for each replication, then 95% of those *intervals* would contain the true mean"
- "Our CI is a range of plausible values for. Values outside the CI are relatively implausible. " (Cumming & Finch, 2005)
- Not about your beliefs about the population

- In an alternative framework – Bayesian Statistics – there's a related idea called "credible intervals". Credible intervals concern beliefs.

Assignment 5, Question 2c

Exercise 2

[a] Use `zl_exp1a_by_subject` to calculate a point estimate of the mean and a 95% confidence interval for each combination of condition and block number.

[b] Recreate the plot that you made in Assignment 4, Exercise 2d, adding the confidence intervals you just calculated in part a.

[c] Were you able to successfully reproduce the confidence intervals from the paper (Fig. 4A)?

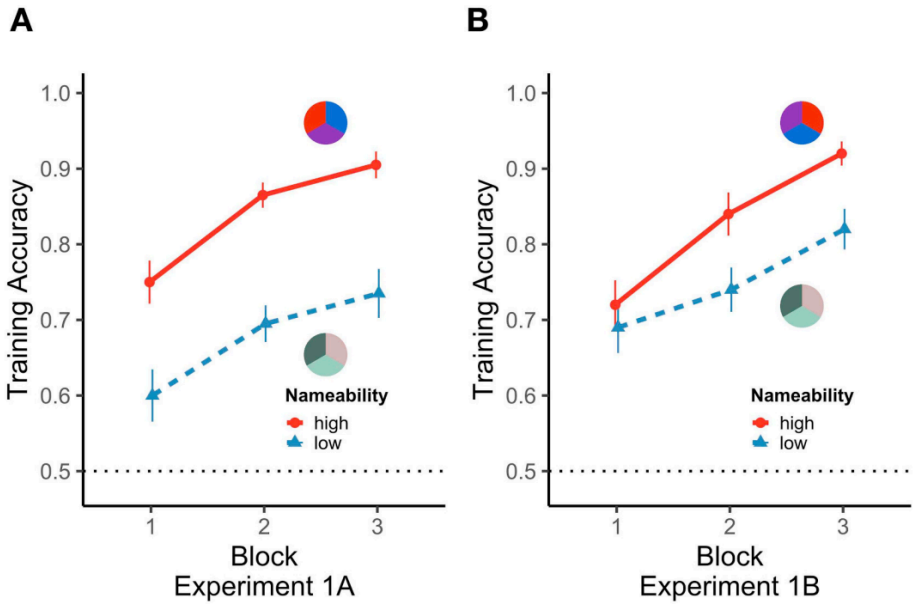


Fig. 4. Accuracy across blocks for (A) Experiment 1A and (B) Experiment 1B. Error bars represent ± 1 SE of the within-subject corrected mean (Morey, 2008).

Midterm review questions (by tomorrow midnight)

Modern Research Methods

Cumulative Science, Big Data and Meta-Analysis

Review Session Questions

In preparation for the review session on Friday, please add two questions/topics you'd like to review in class. The more specific the better (e.g., "how to plot confidence intervals" is better than "confidence intervals").

Last time: Quantifying the magnitude of an effect

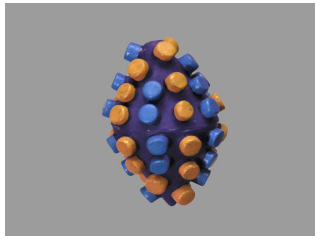
Effect size as unit of analysis: Quantitative, scale-free measure of an effect.

Cohen's *d*:

$$\text{Effect Size} = \frac{\text{diff. between means}}{\text{standard dev.}}$$

Practice with effect sizes

Where's the dofa?



Bion, et al. (2013)

For 24 mo, mean proportion of trials fixating on novel object = .65 (SD = .13)

$$d = \frac{M_{exp} - M_{control}}{SD} = \frac{.65 - .5}{.13} \approx 1.15$$

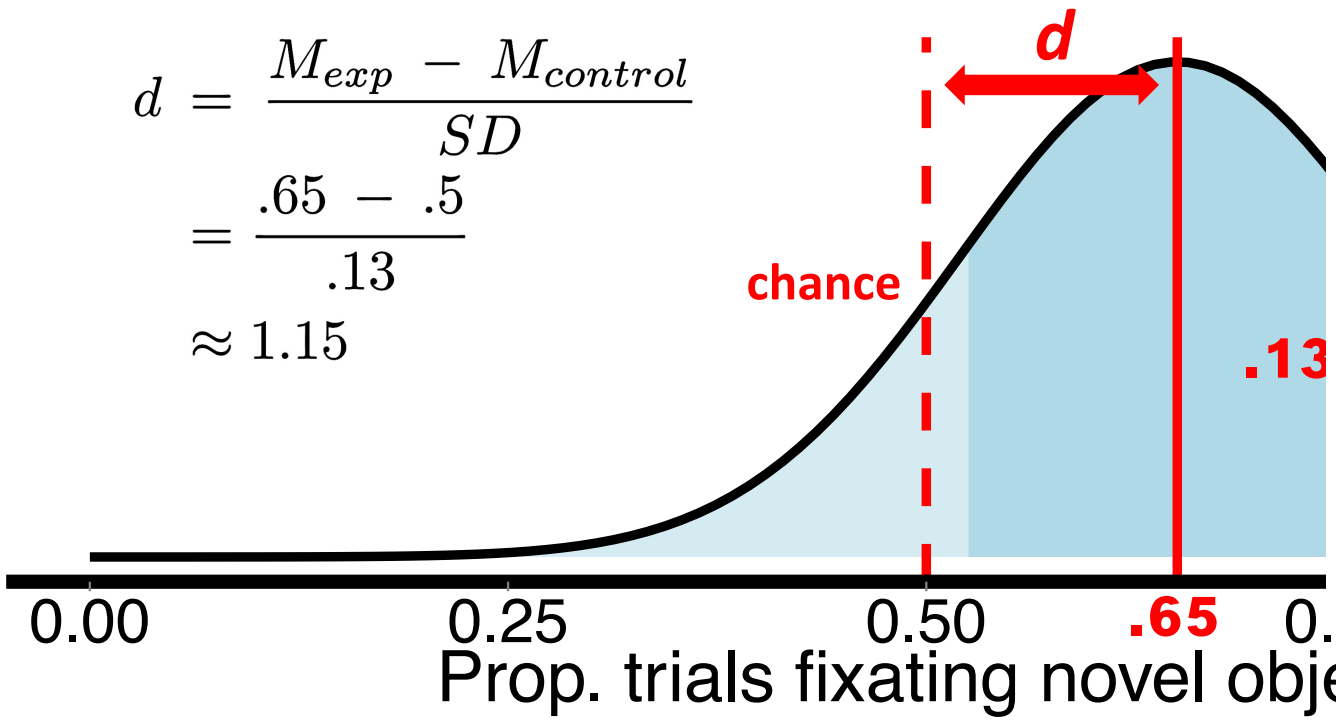
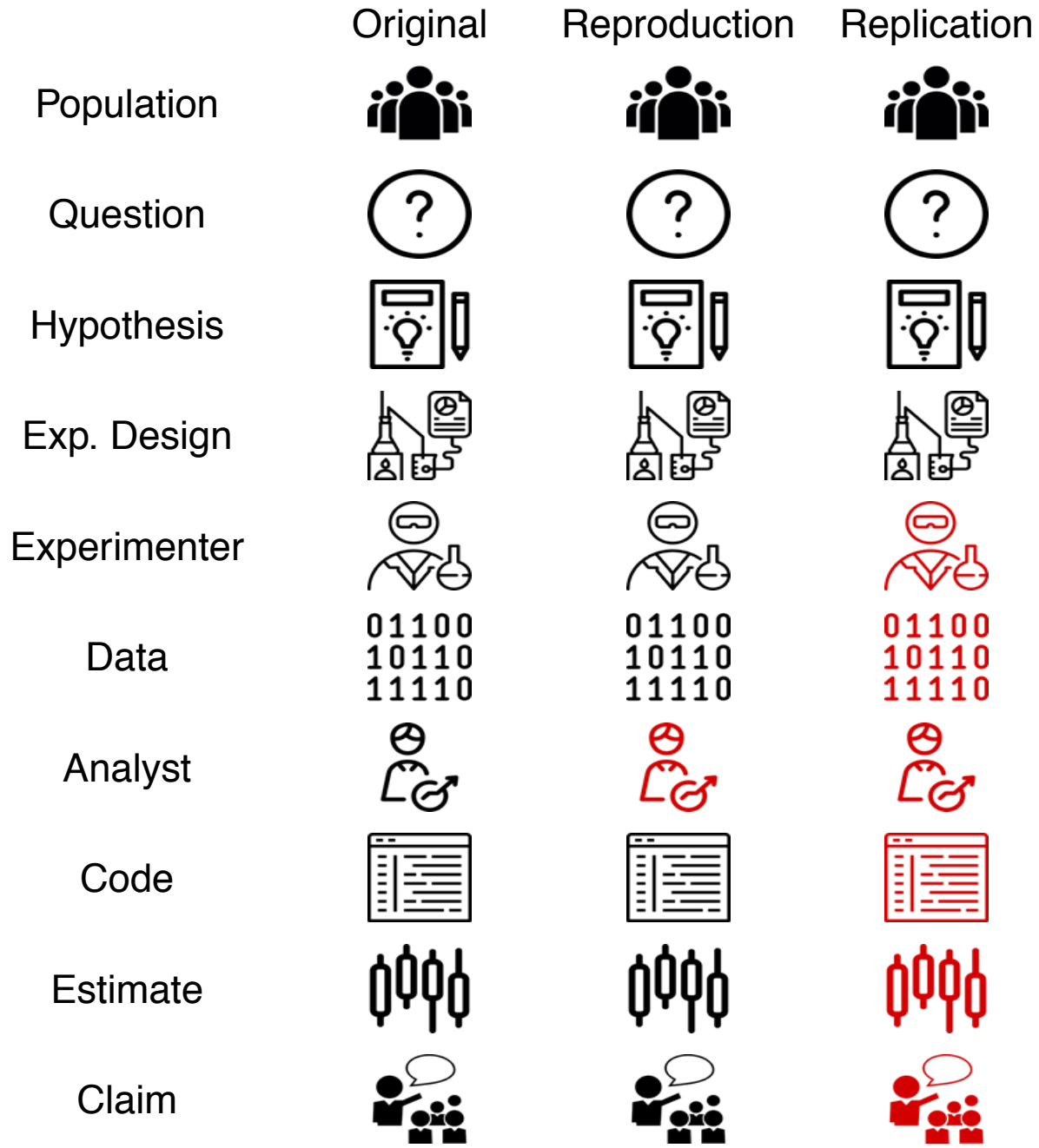


Table 1
Mean, standard deviation, *t* statistics, and *p* value for comparisons against chance (0.50) of accuracy on Familiar-word, Disambiguation, and Retention trials, for the three age groups.

Condition and age	*M*	*SD*	*df*	*t*	*p*
Familiar-word					
18 months	0.66	0.08	21	9.93	0.001
24 months	0.76	0.10	24	12.85	0.001
30 months	0.82	0.08	19	17.67	0.001
Disambiguation					
18 months	0.52	0.14	21	0.59	0.562
24 months	0.65	0.13	24	5.34	0.001
30 months	0.68	0.14	19	5.90	0.001



Original

Different

REPLICATE = Get same result with a new dataset

(Patil, Peng, & Leek, 2019)

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.

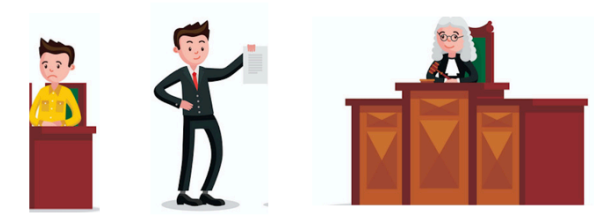
What counts as a replication?

- 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

Statistical framework for thinking about replications

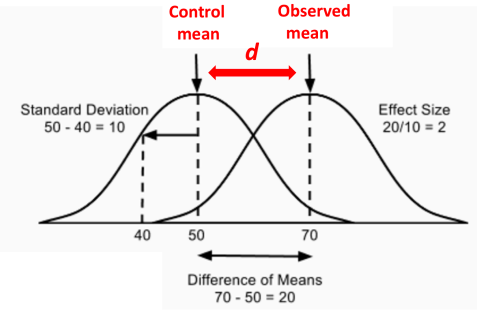
Null Hypothesis Testing

Are the means different? (yes/no)



Effect Sizes

How different are the means?



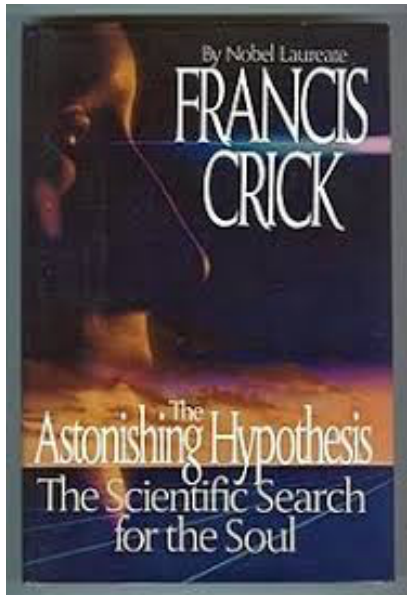
Research Article

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



Read Passage

Anti-free-will essay

Consciousness essay (control)

"glitchy" Math Test

Measure: How much did they cheat?

advocating a deterministic worldview that dismisses individual causation may similarly promote undesirable behavior. In this vein, Peale (1989) bemoaned how quickly and consistently deviant behavior is tagged a "disease," a label that obviates personal responsibility for its occurrence. As a recent *Washington Post* article on neuroscience and moral behavior put it succinctly, "Reducing morality and immorality to brain chemistry—rather than free will—might diminish the importance of personal responsibility" (Vedantam, 2007, p. A01).

Although some people have speculated about the societal risks that might result from adopting a viewpoint that denies personal responsibility for actions, this hypothesis has not been explored empirically. In the two experiments reported here, we manipulated beliefs related to free will and measured their influence on morality as manifested in cheating behavior. We hypothesized that participants induced to believe that human behavior is under the control of predetermined forces would cheat more than would participants not led to believe that behavior is predetermined. Our experimental results supported this hypothesis.

EXPERIMENT 1

Method

Participants

Participants were 30 undergraduates (13 females, 17 males).

Procedure

Participants came to the lab individually. First, according to the condition to which they were randomly assigned, they read one of two passages from *The Astonishing Hypothesis*, a book written by Francis Crick (1994), the Nobel-prize-winning scientist. In the *anti-free-will* condition, participants read statements claiming that rational, high-minded people—including, according to Crick, most scientists—now recognize that actual free will is an illusion, and also claiming that the idea of free will is a side effect of the architecture of the mind. In the *control* condition, participants read a passage from a chapter on consciousness, which did not discuss free will. After reading their assigned material, participants completed the Free Will and Determinism scale (FWD; Paulhus & Margesson, 1994) and the Positive and Negative Affectivity Schedule (PANAS; Watson, Clark, & Tellegen, 1988), which we used to assess whether the reading manipulation affected their beliefs and mood.

Subsequently, participants were given a computer-based mental-arithmetic task (von Hippel, Lakin, & Shakarchi, 2005) in which they were asked to calculate the answers to 20 problems (e.g., $1 + 8 + 18 - 12 + 19 - 7 + 17 - 2 + 8 - 4 = ?$), presented individually. They were told that the computer had a programming glitch and the correct answer would appear on the screen while they were attempting to solve each problem, but that they could stop the answer from being displayed by pressing the space bar after the problem appeared. Furthermore, participants were told that although the experimenter would not know whether they had pressed the space bar, they should try to solve the problems honestly, on their own. In actuality, the computer had been rigged not only to show the answers, but also to record the number of space-bar presses. The dependent measure of cheating was the number of times participants pressed the space bar to prevent the answer from appearing. Afterward, participants were debriefed and thanked for their participation.

Results

Scores on the FWD Scale

We first checked to see whether participants' beliefs about free will were affected by the excerpts they read (anti-free-will vs. control condition). As expected, scores on the Free Will subscale of the FWD scale showed that participants in the anti-free-will condition reported weaker free-will beliefs ($M = 13.6$, $SD = 2.66$) than participants in the control condition ($M = 16.8$, $SD = 2.67$), $t(28) = 3.28$, $p < .01$. Scores on the other three subscales of the FWD scale (Fate, Scientific Causation, and Chance) did not differ as a function of condition, $ts < 1$.

Cheating

We first recoded the dependent measure by subtracting the number of space-bar presses from 20, so that higher scores indicated more cheating. Analysis of the main dependent measure, degree of cheating, revealed that, as predicted, participants cheated more frequently after reading the anti-free-will essay ($M = 14.00$, $SD = 4.17$) than after reading the control essay ($M = 9.67$, $SD = 5.58$), $t(28) = 3.04$, $p < .01$.

Does Rejecting the Idea of Free Will Lead to Cheating?

To test our hypothesis that cheating would increase after participants were persuaded that free will does not exist, we first calculated the correlation between scores on the Free Will subscale and cheating behavior. As expected, we found a strong negative relationship, $r(30) = -.53$, such that weaker endorsement of the notion that personal behavior is determined by one's own will was associated with more instances of cheating.

We next performed a mediation analysis to test our prediction that degree of belief in free will would determine degree of cheating. Using analysis of covariance (ANCOVA), we found support for this hypothesis: When Free Will subscale scores were entered as a predictor of cheating alongside experimental condition, the effect of condition failed to predict cheating, $F < 1$, whereas the effect of free-will beliefs remained significant, $F(1, 27) = 7.81$, $p < .01$.

Ancillary Measure: Mood

To ensure that the essays did not inadvertently alter participants' moods, we assessed positive and negative emotions using the PANAS. Mood did not differ between conditions, $ts < 1.35$, $ps > .19$.

Cheating

We first recoded the dependent measure by subtracting the number of space-bar presses from 20, so that higher scores indicated more cheating. Analysis of the main dependent measure, degree of cheating, revealed that, as predicted, participants cheated more frequently after reading the anti-free-will essay ($M = 14.00$, $SD = 4.17$) than after reading the control essay ($M = 9.67$, $SD = 5.58$), $t(28) = 3.04$, $p < .01$.

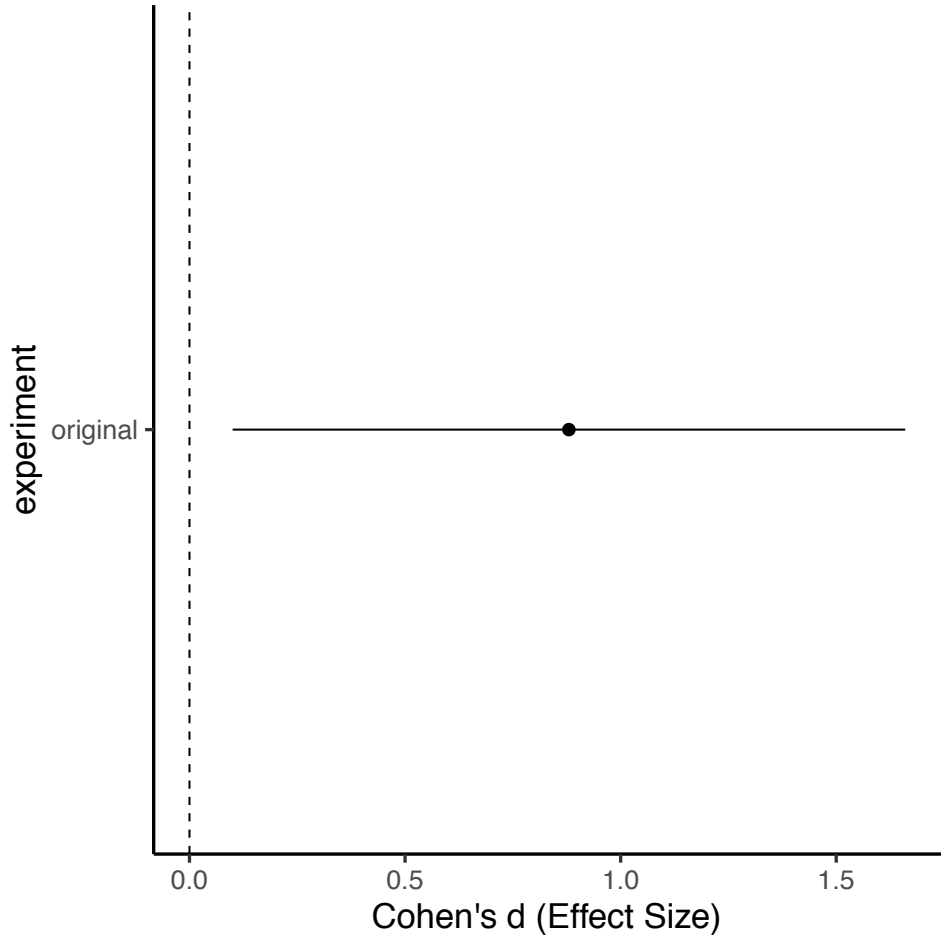
Calculating an effect size

$$\text{Effect Size} = \frac{\text{diff. between means}}{\text{standard dev.}}$$

Cheating

We first recoded the dependent measure by subtracting the number of space-bar presses from 20, so that higher scores indicated more cheating. Analysis of the main dependent measure, degree of cheating, revealed that, as predicted, participants cheated more frequently after reading the anti-free-will essay ($M = 14.00$, $SD = 4.17$) than after reading the control essay ($M = 9.67$, $SD = 5.58$), $t(28) = 3.04$, $p < .01$.

```
> mes(14, 9.67, 4.17, 5.58, 15, 15, verbose = F) %>%
+   select(d, l.d, u.d) %>%
+   rename(lower_ci = l.d,
+          upper_ci = u.d)
     d lower_ci upper_ci
1 0.88      0.1     1.66
```

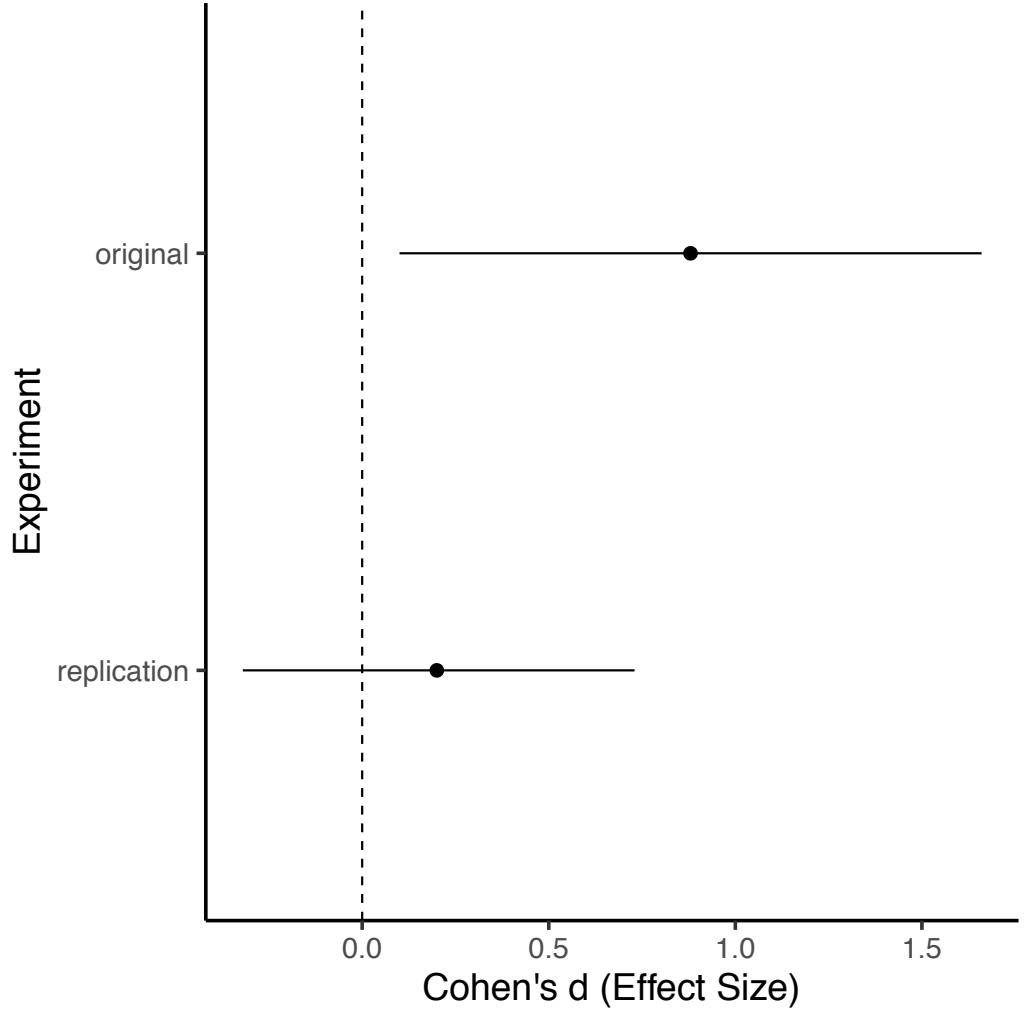


Replication of Vohs and Schooler (2017)

$N = 58$

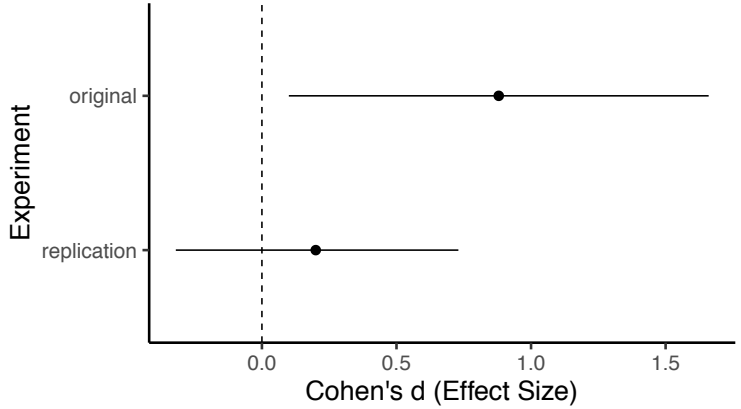
```
> t.test(determinism, freewill)

        Welch Two Sample t-test

data:  determinism and freewill
t = 0.77137, df = 50.951, p-value = 0.4441
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.934273  4.348066
sample estimates:
mean of x mean of y
 6.793103  5.586207
```

```
> mes(6.793103,  5.586207,  6.831541, 4.931801, 29, 29,
+     verbose = F) %>%
+   select(d, l.d, u.d) %>%
+   rename(ci_lower = l.d,
+          ci_upper = u.d)
    d ci_lower ci_upper
1 0.2    -0.32     0.73
```



Interpreting the replication

Statistics	Original	Replication	Interpretation
p-value (*t*-test)	<.01	.44	Did not replicate
Effect size	.84 [.06, 1.62]	.2 [-.32, .72]	Effect size much smaller (1/4), and the confidence interval for the replicates includes 0.



Talk to the person(s) next to you, and generate a list of reasons why the Vohs & Schooler effect might not have replicated.

Why might an effect not replicate?

1. Effect isn't real, got unlucky in original.
2. Effect is real, but got unlucky in replication.

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

How replicable are psychological studies?

Rumblings that something isn't right…

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition[1]

Edward Vul,[1] Christine Harris,[2] Piotr Winkielman,[2] & Harold Pashler[2]

[1]*Massachussetts Institute of Technology and* [2]*University of California, San Diego*

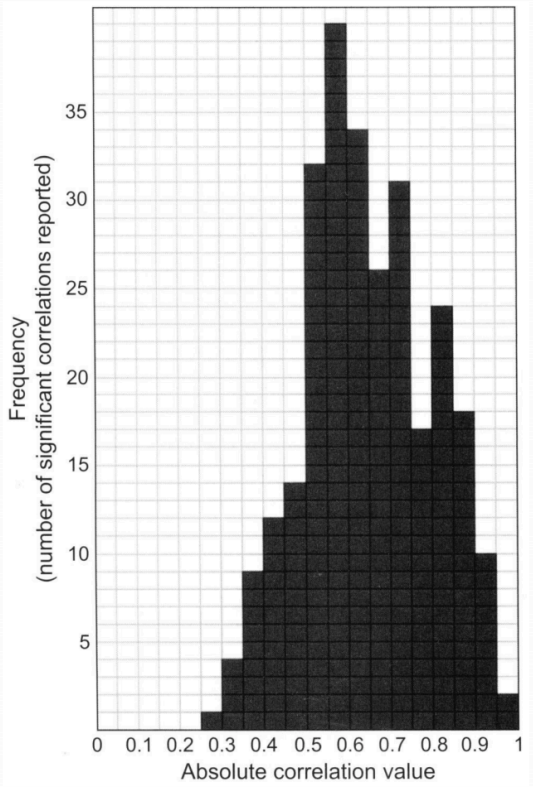


Fig. 1. A histogram of the correlations between evoked blood oxygenation level dependent response and behavioral measures of individual differences seen in the studies identified for analysis in the current article.

PSYCHOLOGY

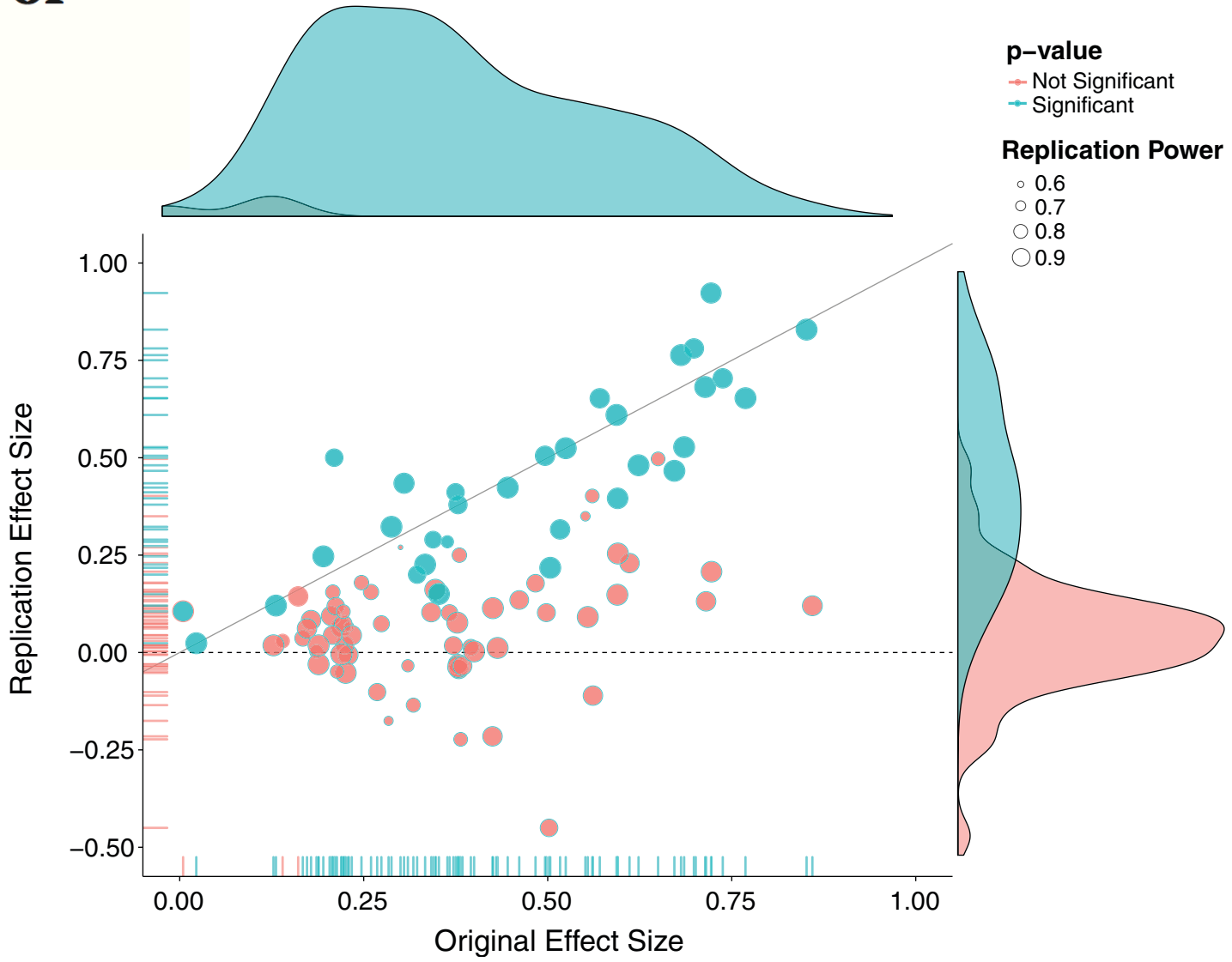
Estimating the reproducibility of psychological science

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)



Why might an effect not replicate?

1. **Effect isn't real, got unlucky in original.**
2. Effect is real, but got unlucky in replication.

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

- More than 5% of studies are failing to replicate – why?
- There must be other reasons that researchers are getting a positive effect, when there is no effect.

REPLICATION CRISIS

- Sometimes known as the "credibility crisis"
- Centered on, but not limited to, social psychology
- A period, from ~2011 to now, of intense worry about the credibility of research
- A big part of the credibility crisis has been figuring out what's not working

Slide adapted from Patrick S. Forscher

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 # 1: Fraud (i.e, researchers are just making up their data)

Harvard Dean Confirms Misconduct in Hauser Investigation

by Greg Miller on 20 August 2010, 3:11 PM | 2 Comments

Email Print | +1 0 More PREVIOUS ARTICLE

In an e-mail sent earlier today to Harvard University faculty members, Michael Smith, dean and Sciences (FAS), confirms that cognitive scientist Marc Hauser "was found solely respo thorough investigation by a faculty member investigating committee, for eight instances of s under FAS standards."

BuzzFeed News

REPORTING TO YOU

SIGN IN ABOUT US GOT A TIP? SUPPORT US BUZZFEED.COM SECTIONS

y Petito Southwest Airlines Delays War Crimes Boston Marathon Remote Learning Democrats' Social Spending Bill

SCIENCE

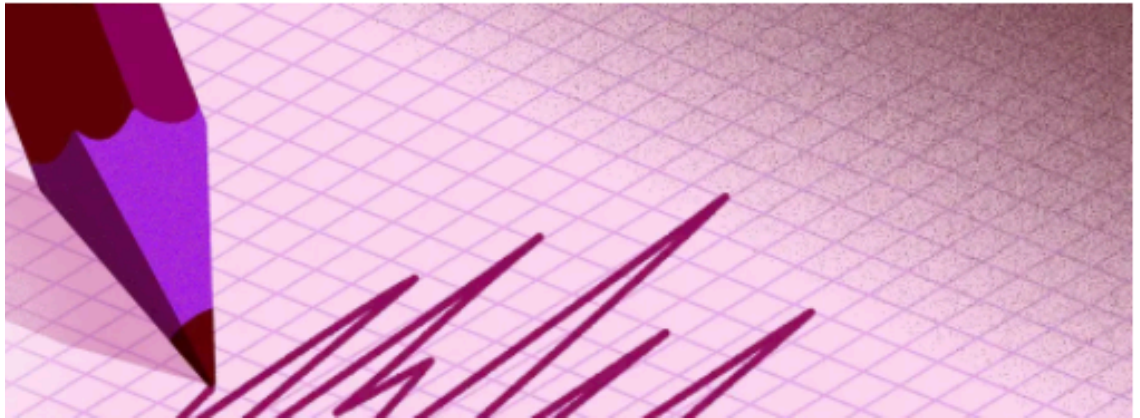
A Famous Honesty Researcher Is Retracting A Study Over Fake Data

Renowned psychologist Dan Ariely literally wrote the book on dishonesty. Now some are questioning whether the scientist himself is being dishonest.

Stephanie M. Lee
BuzzFeed News Reporter

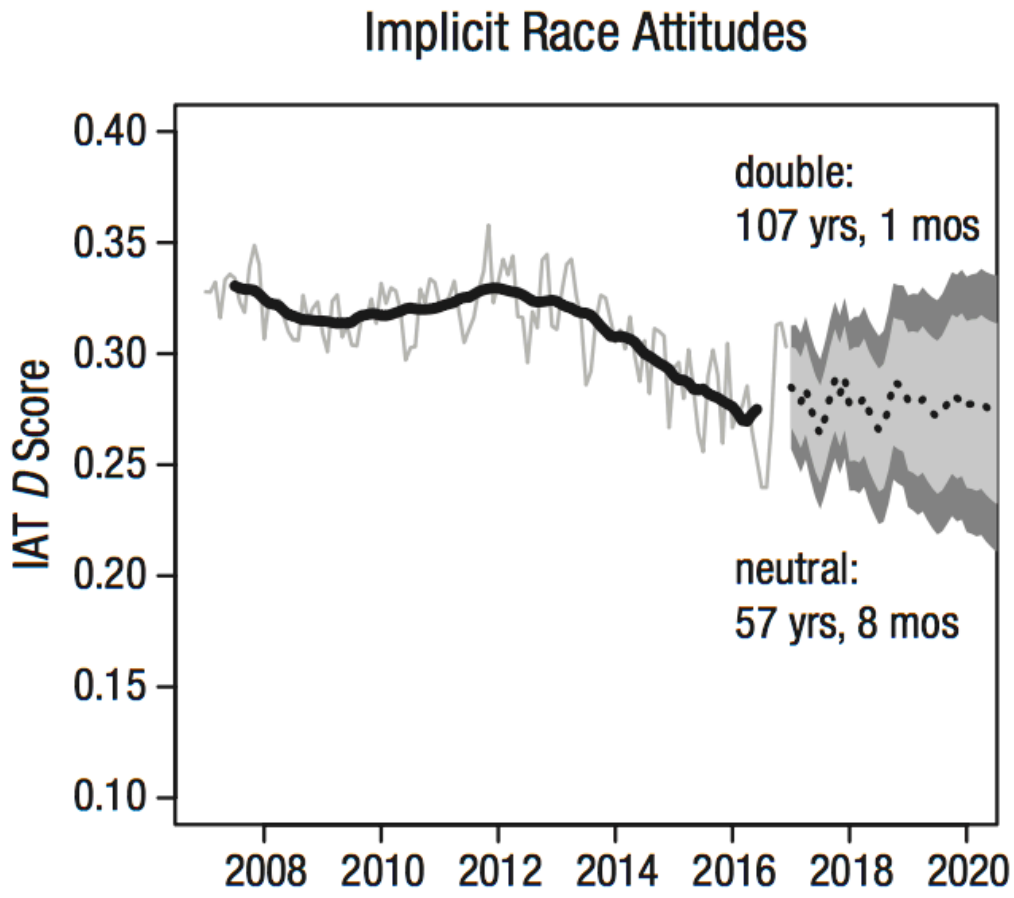
Last updated on August 25, 2021, at 1:15 p.m. ET
Posted on August 20, 2021, at 2:40 p.m. ET

Tweet Share Copy



This is likely very rare!

Reason #2: Actual change in population effect



(Charlesworth & Banaji, 2019)

Next Time: More sources of 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