False-Positives, p-Hacking, Statistical Power, and Evidential Value

Leif D. Nelson

University of California, Berkeley

Haas School of Business

Summer Institute

June 2014

BITSS

Berkeley Initiative for Transparency in the Social Sciences

Who am I?

- Experimental psychologist who studies judgment and decision making.
  - And has interests in methodological issues

Who are you?

[*not a rhetorical question*]

- Grad Student vs. Post-Doc vs. Faculty?
- Psychology vs. Economics vs. Other?
- Have you read any papers that I have written?
  - Really? Which ones?

Things I want you to get out of this

- It is quite easy to get a false-positive finding through p-hacking. (5%)
- Transparent reporting is critical to improving scientific value. (5%)
- It is (very) hard to know how to correctly power studies, but there is no such thing as overpowering. (30%)
- You can learn a lot from a few p-values. (remainder %)

This will be most helpful to you if you ask questions.

A discussion will be more interesting than a lecture.

SLIDES ABOUT P-HACKING

False-Positives are Easy

- It is common practice in all sciences to report less than everything.
  - So people only report the good stuff. We call this *p*-Hacking.
  - Accordingly, what we see is too “good” to be true.
  - We identify six ways in which people do that.

Six Ways to p-Hack

1. Stop collecting data once p<.05
2. Analyze many measures, but report only those with p<.05.
3. Collect and analyze many conditions, but only report those with p<.05.
4. Use covariates to get p<.05.
5. Exclude participants to get p<.05.
6. Transform the data to get p<.05.

OK, but does that matter very much?

- As a field we have agreed on $p<.05$. (i.e., a 5% false positive rate).
- If we allow p-hacking, then that false positive rate is actually 61%.
- Conclusion: p-hacking is a potential catastrophe to scientific inference.

P-Hacking is Solved Through Transparent Reporting

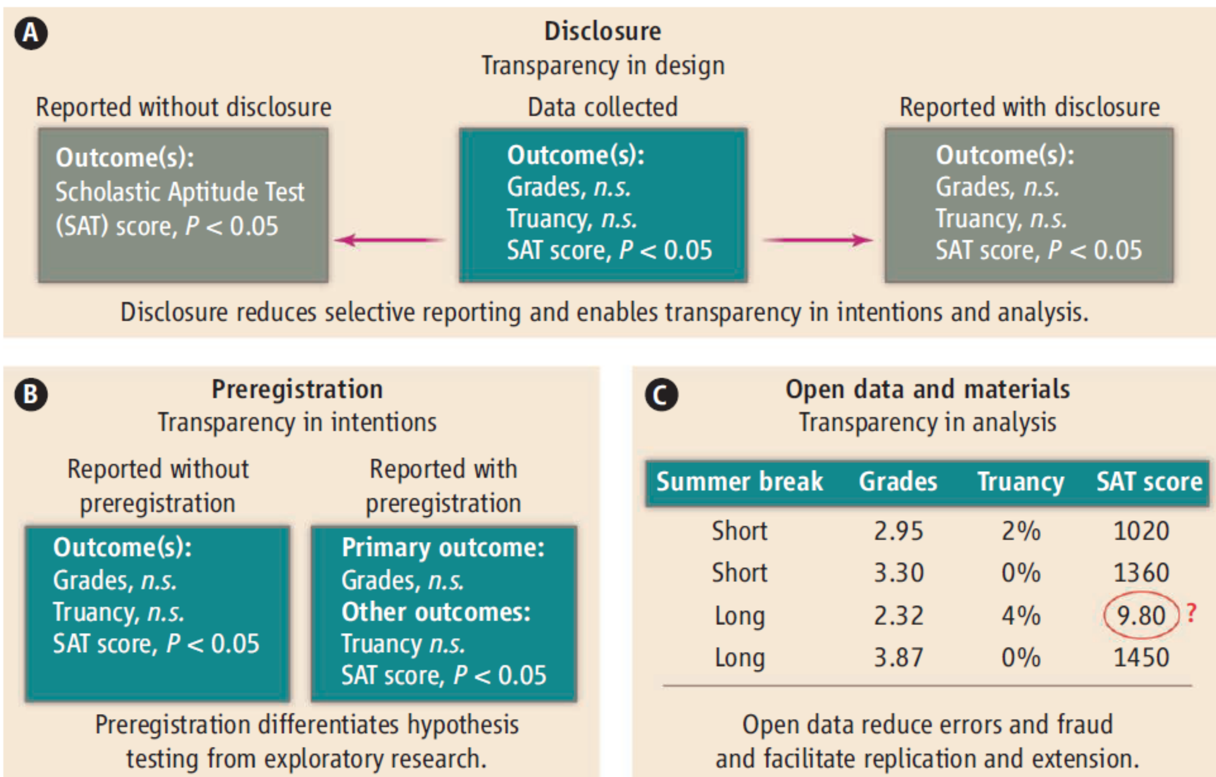
- Instead of reporting only the good stuff, just report all the stuff.

P-Hacking is Solved Through Transparent Reporting

- Solution 1:
  1. Report sample size determination.
  2. N>20 [note: I will tell you later about how this number is insanely low. Sorry. Our mistake.]
  3. List all of your measures.
  4. List all of your conditions.
  5. If excluding, report without exclusion as well.
  6. If covariates, report without.

P-Hacking is Solved Through Transparent Reporting

- Solution 2:



Three mechanisms for increasing transparency in scientific reporting. Demonstrated with a research question: "Do shorter summer breaks improve educational outcomes?" n.s. denotes $P > 0.05$.

P-Hacking is Solved Through Transparent Reporting

- Implications:
  - Exploration is necessary; therefore replication is as well.
  - Without p-hacking, fewer significant findings; therefore fewer papers.
  - Without p-hacking, need more power; therefore more participants.

SLIDES ABOUT POWER

Motivation

- With *p*-hacking,
  - statistical power is irrelevant, most studies work
- Without *p*-hacking.
  - take power seriously, or most studies fail
- Reminder. Power analysis:
  - **Guess** effect size (d)
  - **Set** sample size (n)
- Our question: Can we make guessing *d* easier?
- Our answer: **No**
- Power analysis is not a practical way to take power seriously

How to guess d?

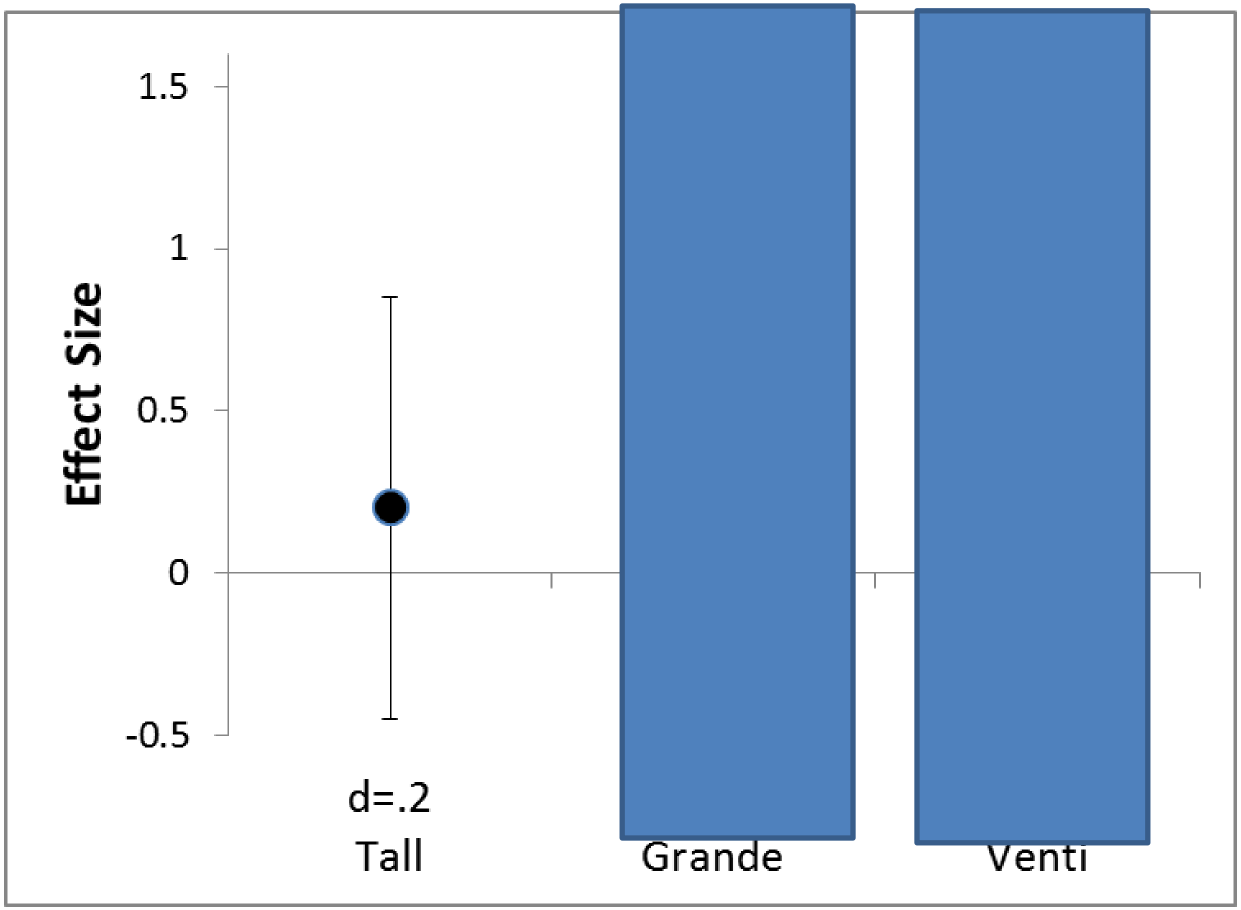
- Pilot
- Prior literature
- Theory/gut

Some kind words before the bashing

- Pilots: They are good for:
  - Do participants get it?
  - Ceiling effects?
  - Smooth procedure?
- Kind words end here.

Pilots: useless to set sample size

- Say Pilot: n=20
  - $\hat{d}$ = .2
  - $\hat{d}$ = .5
  - $\hat{d}$ = .8



- **In words**
  - Estimates of *d* have too much sampling error.

- **In more interesting words**
  - Next.

Think of it this way

Say in actuality you need *n*=75

Run Pilot: n=20

What will Pilot say you need?

- Pilot 1: “you need *n*=832”
- Pilot 2: “you need *n*=53”
- Pilot 3: “you need *n*=96”
- Pilot 4: “you need *n*=48”
- Pilot 5: “you need *n*=196”
- Pilot 6: “you need *n*=10”
- Pilot 7: “you need *n*=311”

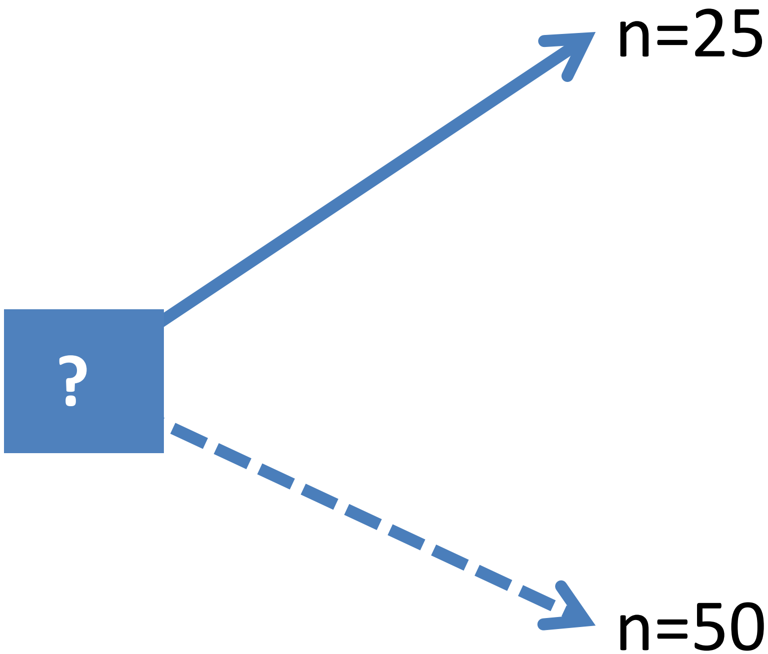
Thanks Pilot!

n=20 is not enough.

How many subjects do you need

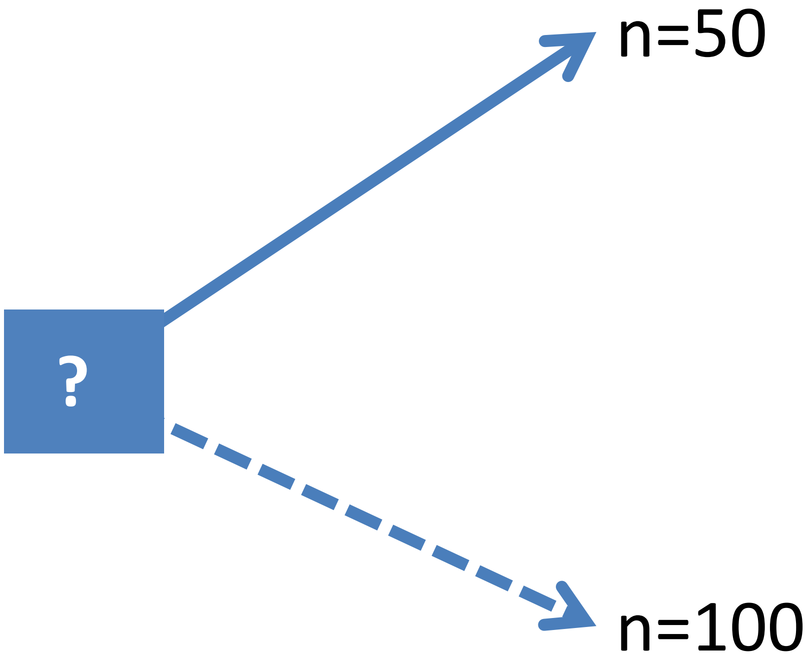
to know

how many subjects you need?



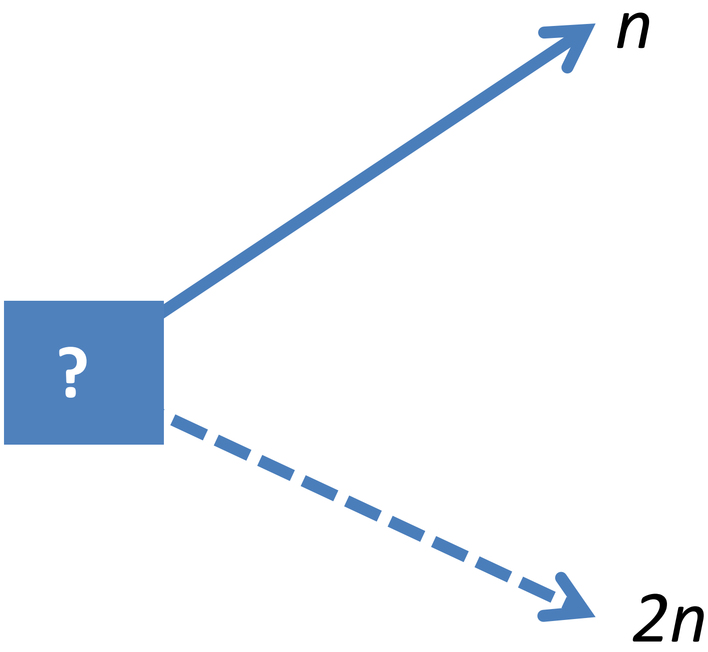
Need a Pilot with…

n=133



Need a Pilot with…
n=276

“Theorem” 1



Need: 5n

How to guess d?

- Pilot
- Existing findings
- Theory/gut

Existing findings

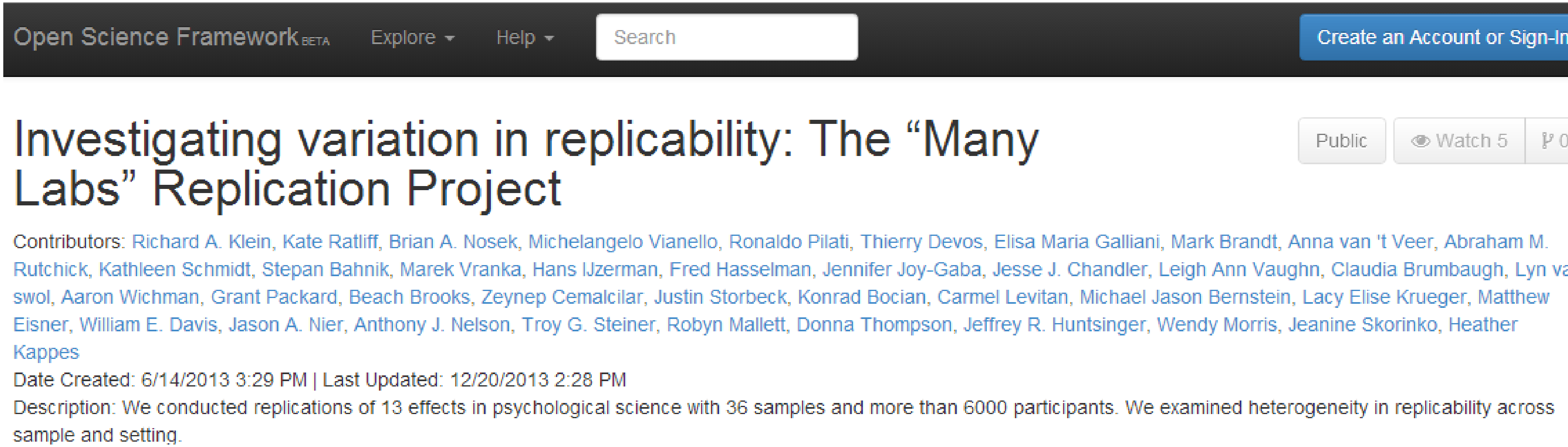
- **One hand**
  - Larger samples
- **Other hand**
  - Publication bias
  - More noise
    - $\neq$ sample
    - $\neq$ design
    - $\neq$ measures

Best (im)possible case scenario

- Would guessing d be reasonable based on other studies?

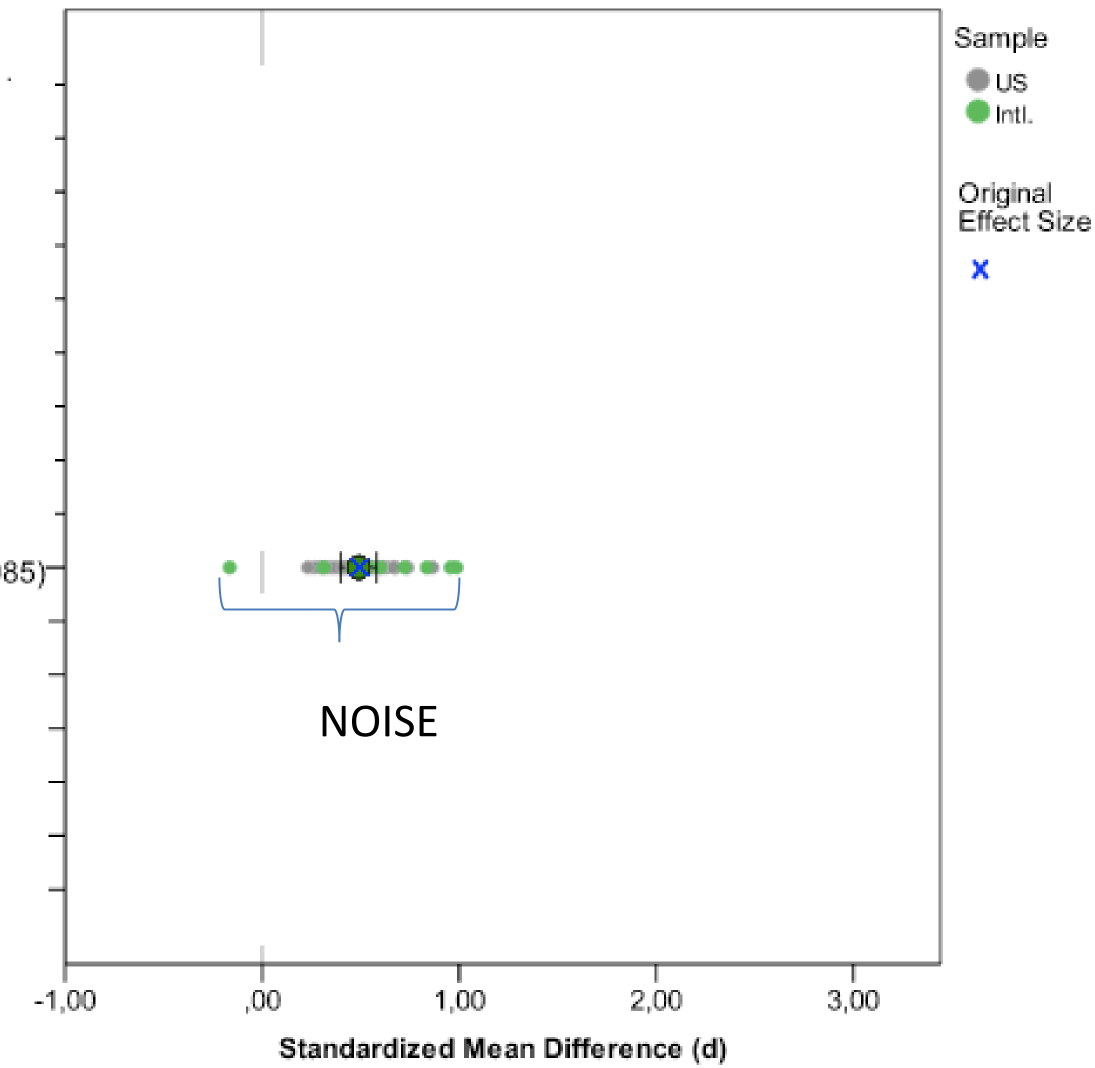
“Many Labs” Replication Project

- Klein et al.,
- 36 labs
- 12 countries
- N=6344
- Same 13 experiments



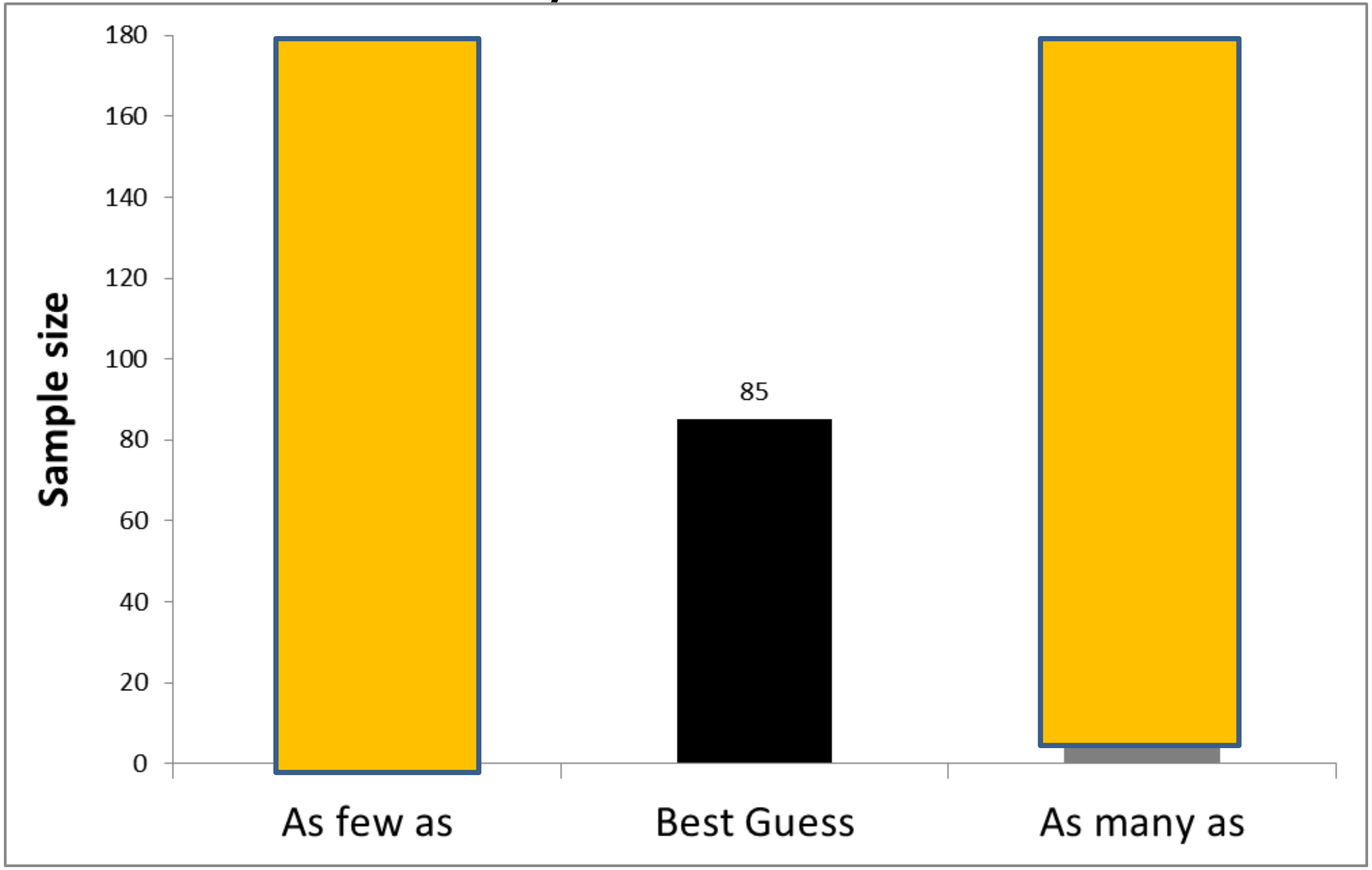
How much TV per day?

Low-frequency alternatives	High-frequency alternatives
Up to ½ hour	Up to 2½ hours
½ hour to 1 hour	2½ hours to 3 hours
1 hour to 1½ hours	3 hours to 3½ hours
1½ hours to 2 hours	3½ hours to 4 hours
2 hours to 2½ hours	4 hours to 4½ hours
More than 2½ hours	More than 4½ hours



If 5 identical studies already done

- Best guess: n=85
- How sure are you?



Best case scenario gives range 3:1

Reality is massively worse

- Nobody runs 6th identical study.
  - Moderator: Fluency
  - Mediator: Perceived-norms
  - DV: 'Real' behavior
- Publication bias

Where to get *d* from?

- Pilot
- Existing findings
- $\boxed{\text{Theory/gut}}$

Say you think/feel d~.4

d=.44 ~ .4

→n=83

d=.35, ~ .4

→n=130

Rounding error → 100 more participants

Transition (key) slide

- Guessing d is completely impractical

  → Power analysis is also.
- Step back: Problem with underpowering?

- Unclear what failure means.

- Well, when you put it that way:

Let's power so that we know what failure means.

Existing view

1. **Goal**: **Success**

2. **Guess d**

3. **Set n:**

   "80%" success

New View

1. **Goal**: Learn from results

2. **Accept d is unknown**

   If interesting → 0 possible

   If 0 possible → very small possible

3. **Set n:**

   100% learning

   Works: keep going

   Fails: Go Home

What is “Going *Big*”?

A. Limited resources (most cases)

(e.g., lab studies)

– What *n* are you ***willing to pay*** for this effect?

– Run n

- Fails, too small for me.
- Works, keep going, adjust n.

B. ‘Unlimited’ resources (fewest cases)

(e.g., Project Implicit, Facebook)

– Smallest effect you *care* about

SLIDES ABOUT P-VALUES

Defining Evidential Value

- **Statistical significance**
  Single finding:
  unlikely result of chance

Could be caused by selective reporting rather than chance

- **Evidential value**
  Set of significant findings:
  unlikely result of selective reporting

Motivation: we only publish if p<.05

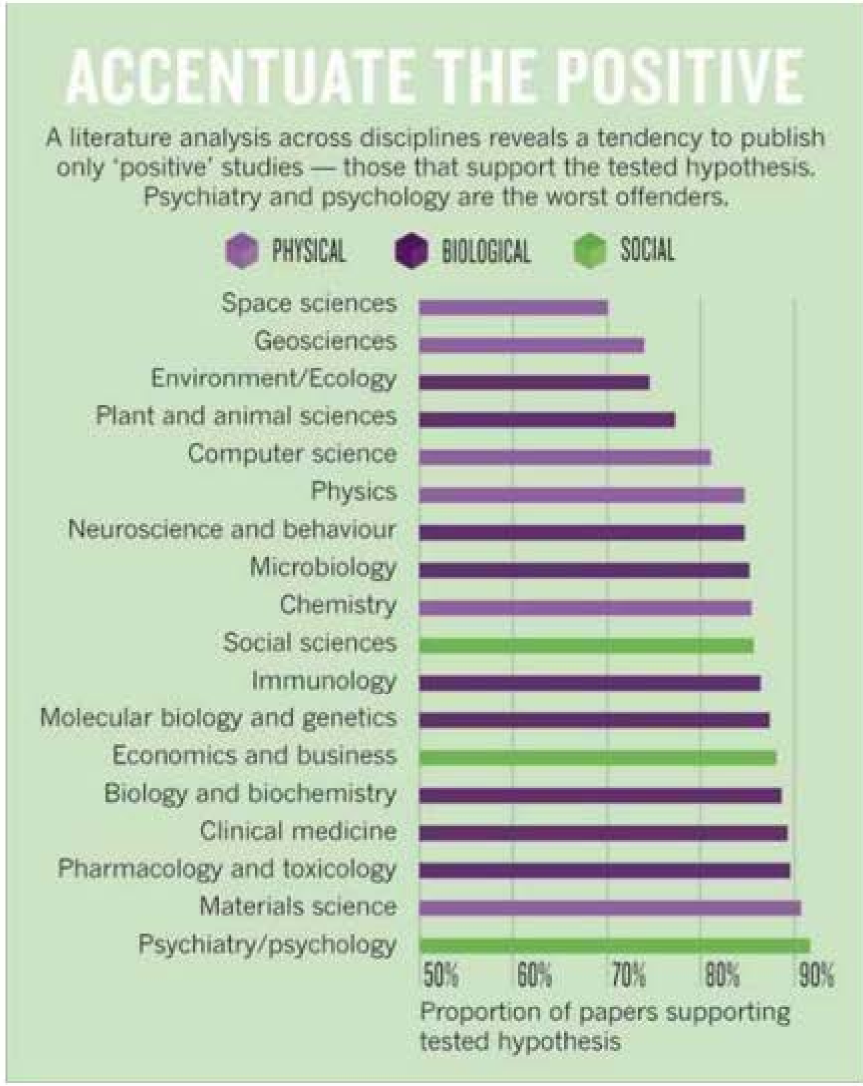


Figure 1: From Fanelli, D. Scientometrics 90, 891–904 (2011).

Motivation

Nonexisting effects: only see false-positive evidence

Existing effects: only see strongest evidence

Published scientific evidence is not representative of reality.

Outline

- Shape
- Inference
- Demonstration
- How often is p-curve wrong?
- Effect size estimation
- Selecting *p*-values

p-curve's shape

- **Effect does not exist:** flat
- **Effect exists**: right-skew.

  (more lows than highs)
- ***Intensely p*-hacked**: left-skew

  (more highs than lows)

Why flat if null is true?

***p*-value:**

prob(result | null is true).

Under the null:

- What percent of findings p ≤.30
  - 30%
- What percent of findings p ≤.05
  - 5%
- What percent of findings p ≤.04
  - 4%
- What percent of findings p ≤.03
  - 3%

Got it.

Why more lows than high if true?

(right skew)

- Height: men vs. women
- *N* = Philadelphia
- What result is more likely?

  In Philadelphia, men taller than women (*p*=.047)

  (p=.007)

- Not into intuition?

  Differential convexity of the density function

  Wallis (Econometrica, 1942)

Why left skew with *p*-hacking?

- Because *p*-hackers have limited ambition
- *p*=.21
  → Drop if >2.5 SD
- *p*=.13
  → Control for gender
- *p*=.04
  → Write Intro

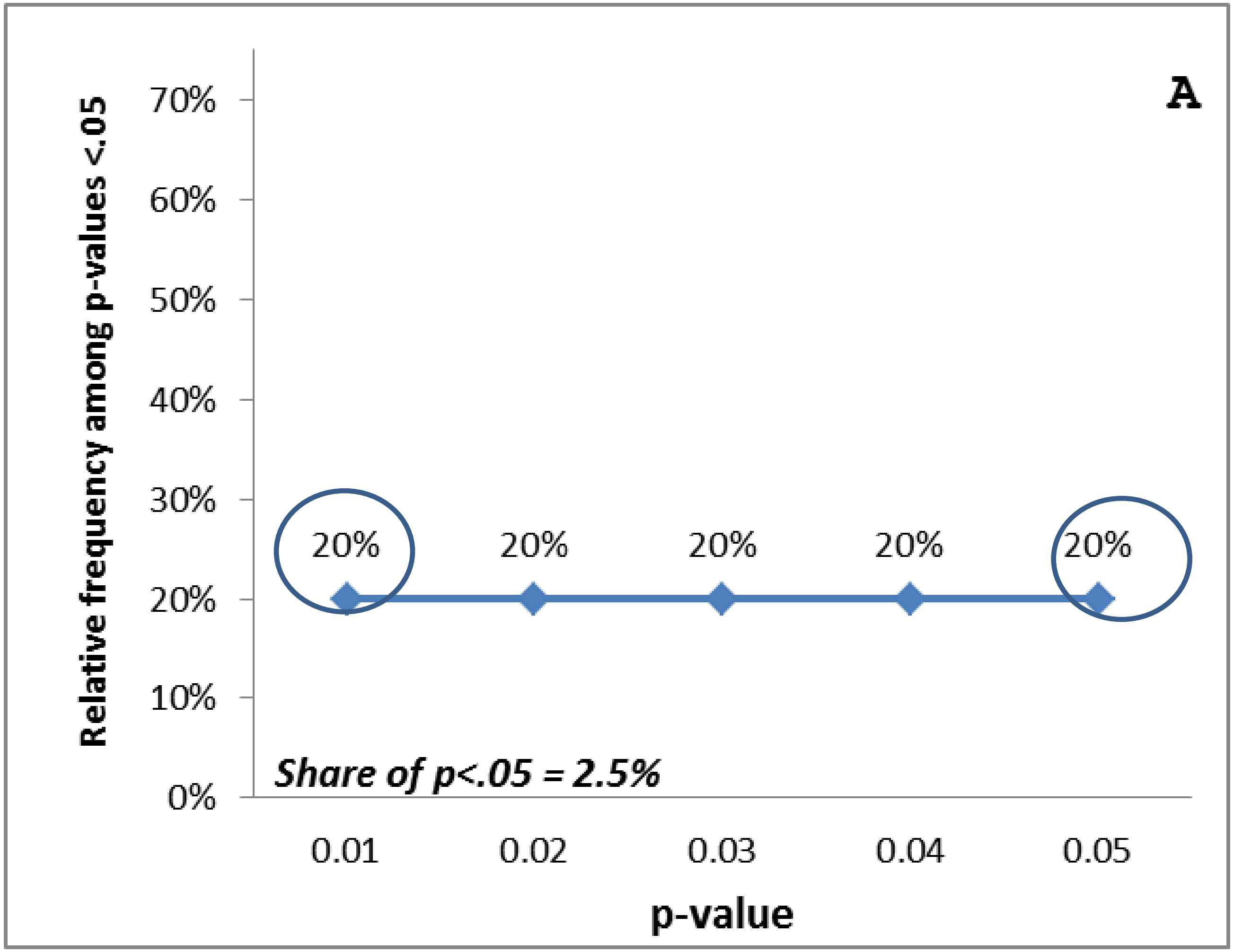
- If we stop p-hacking as soon as *p*<.05,
- Won't get to *p*=.02 very often.

Plotting Expected *P*-curves

- Two-sample *t*-tests.
- True effect sizes
  - d=0, d=.3, d=.6, d=.9
- *p*-hacking
  - No: n=20
  - *Yes:* n={20,25,30,35,40}

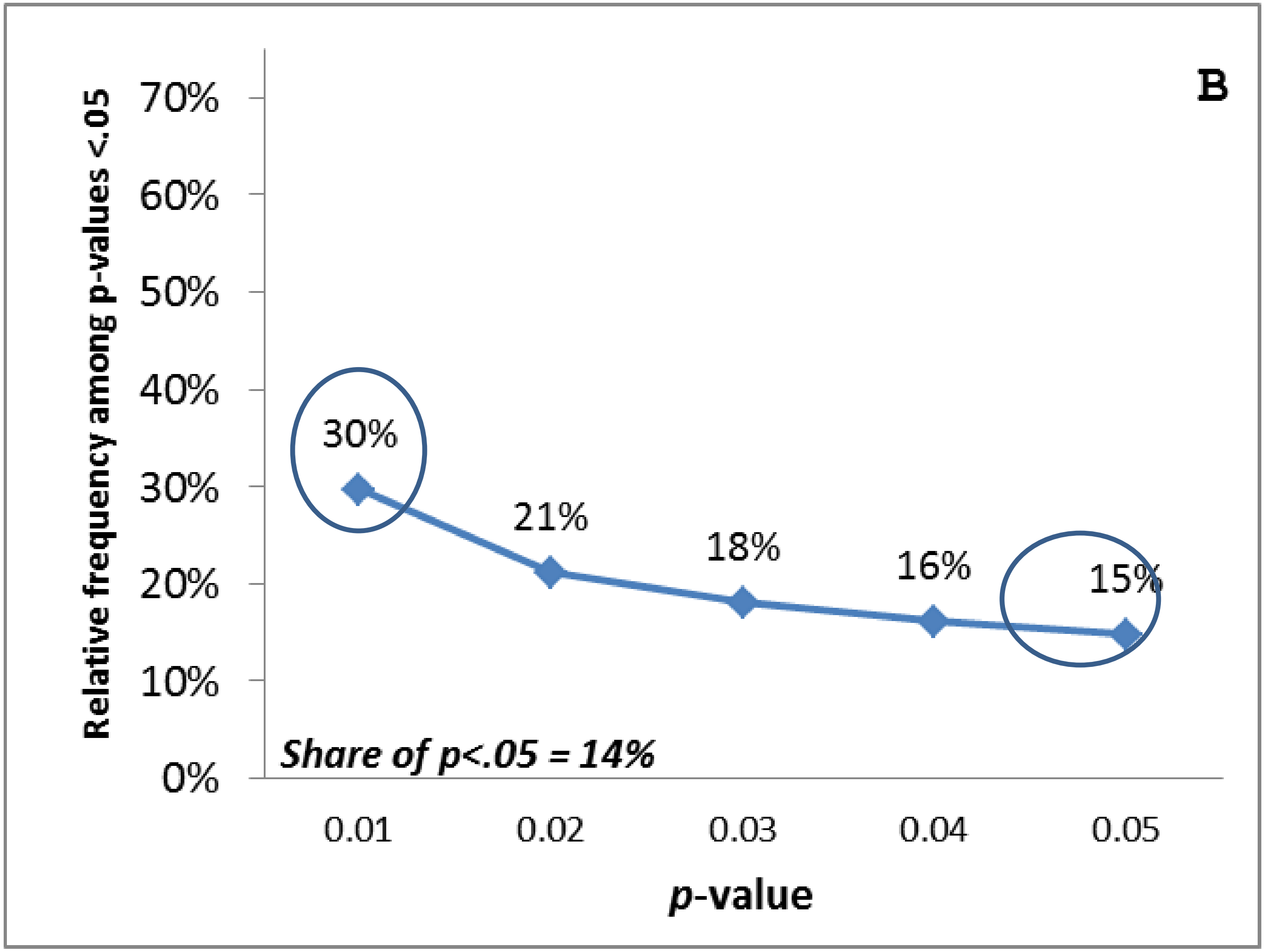
Nonexisting effect (n=20, *d*=0)

As many *p*<.01 as p>.04



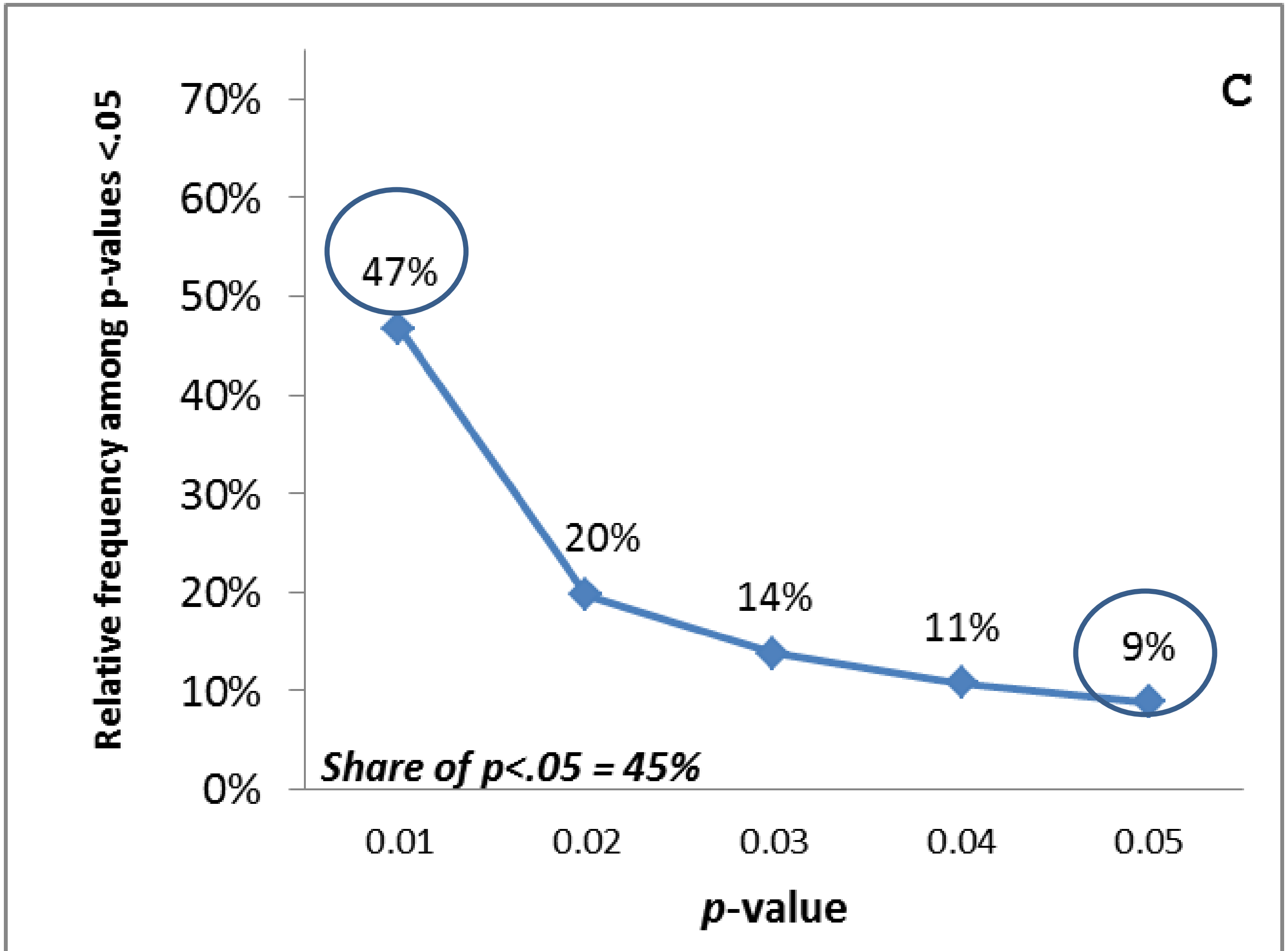
n=20, *d*=.3 / power=14%

Two p<.01 for every p>.04



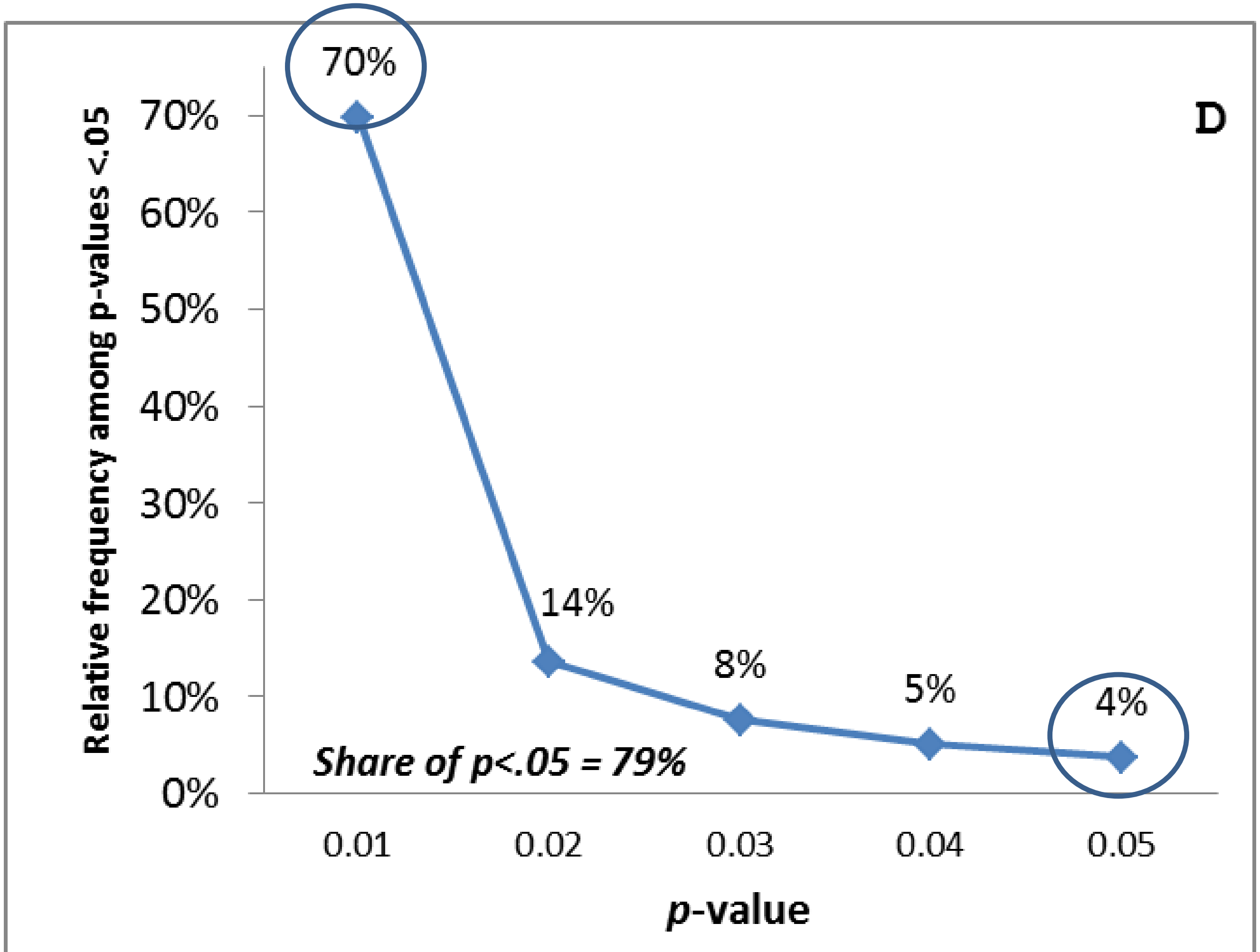
n=20, *d*=.6 / power = 45%

Five *p*<.01 per every one *p*>.04



n=20, *d*=.9 / power=79%

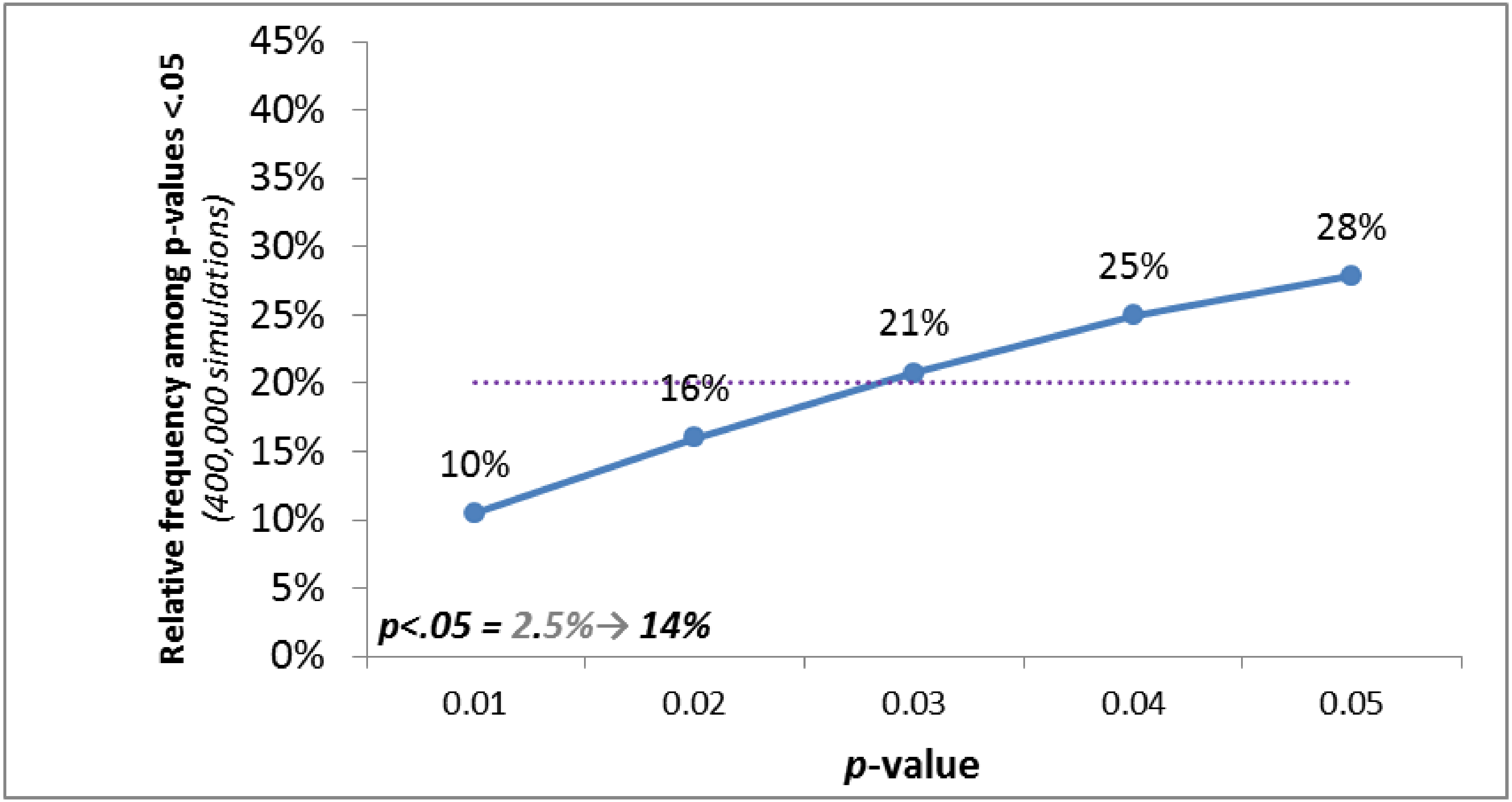
Eigtheen *p*<.01 per every *p*>.04.



Adding *p*-hacking

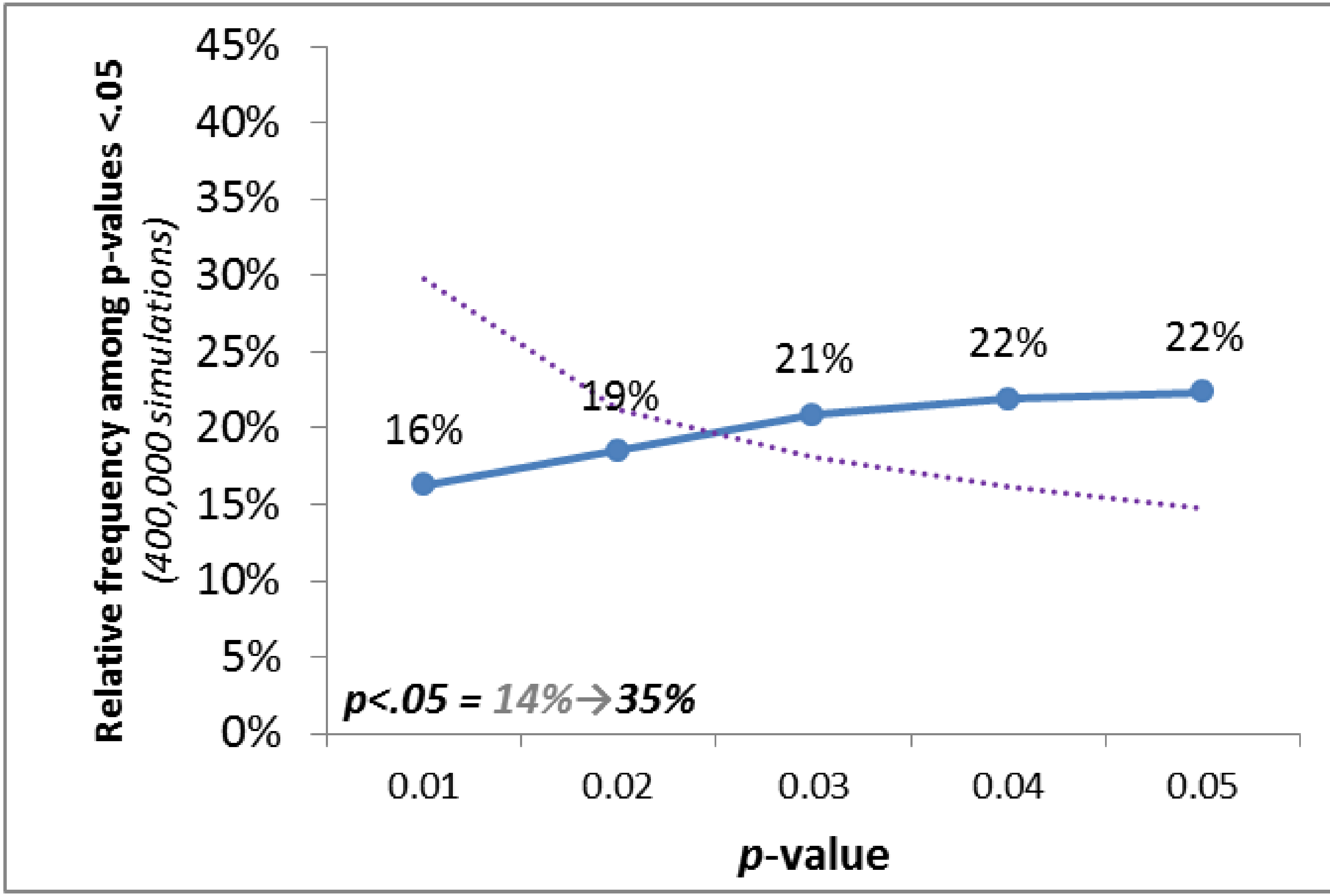
$n=\{20,25,30,35,40\}$

d=0

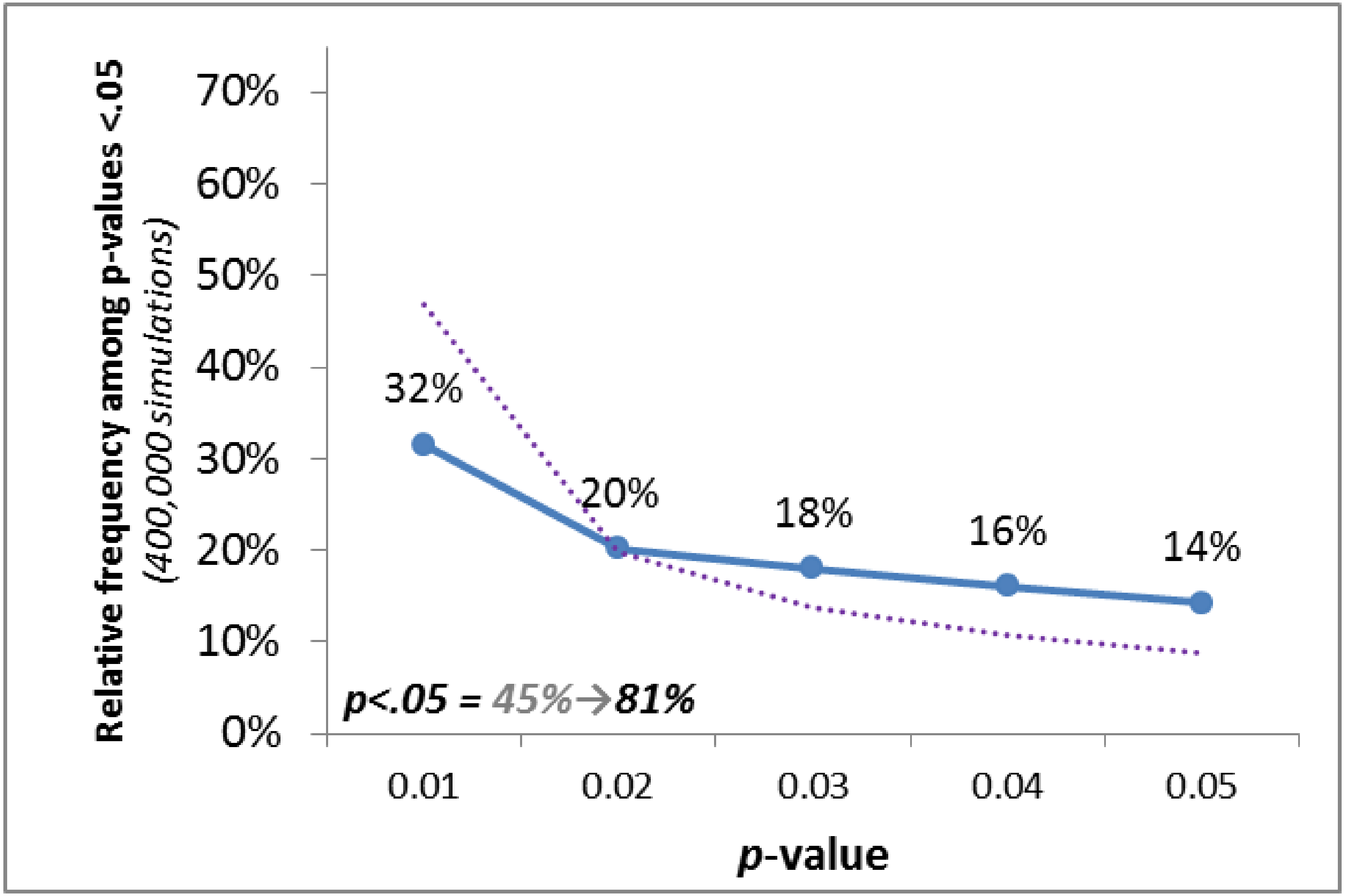


Relative frequency among p-values <.05
(400,000 simulations)
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
10%
16%
21%
25%
28%
p<.05 = 2.5%→ 14%
0.01
0.02
0.03
0.04
0.05
p-value

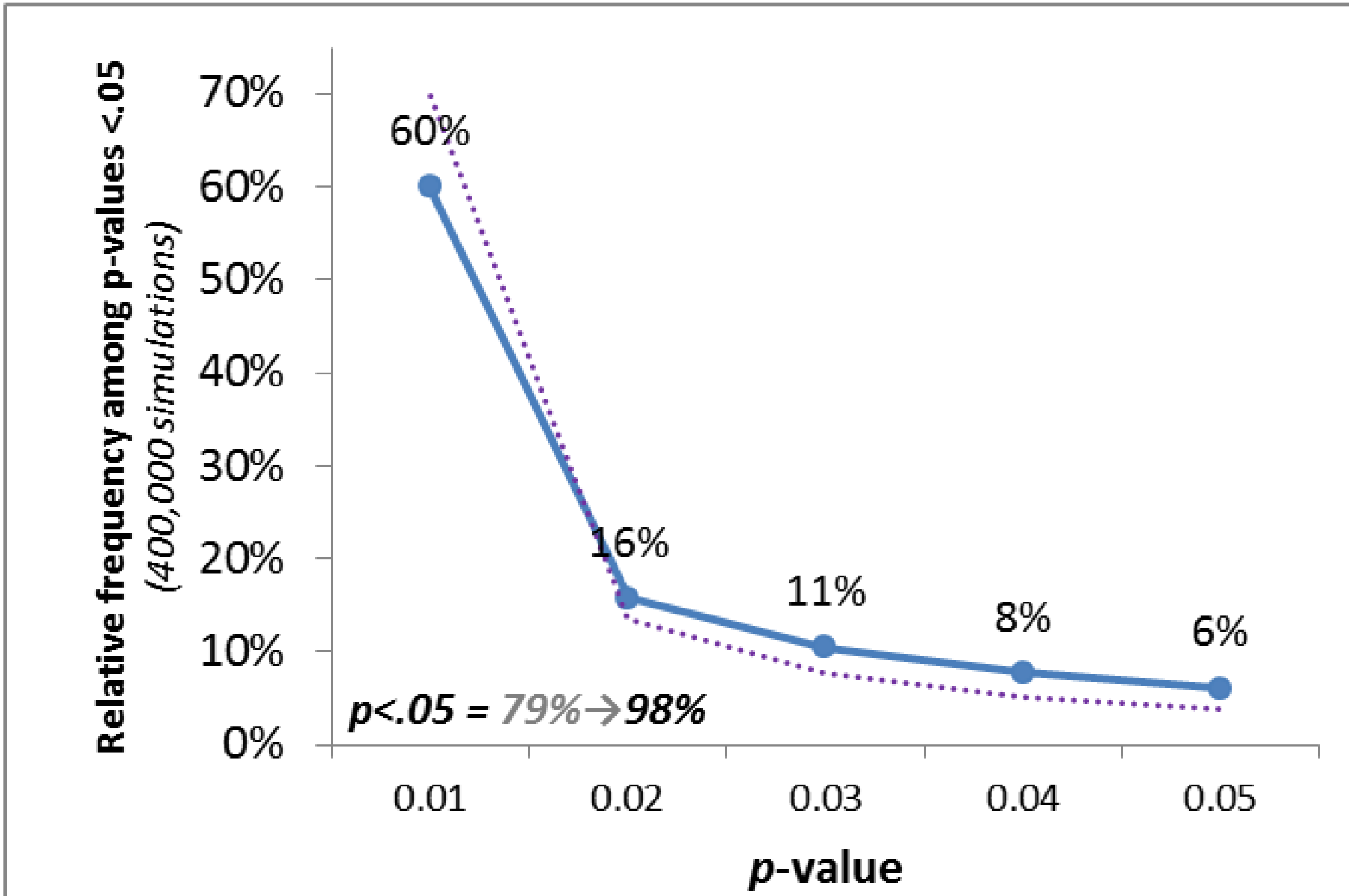
d=.3 / original power=14%



d=.6 / original-power = 45%



d=.9 / original-power=79%

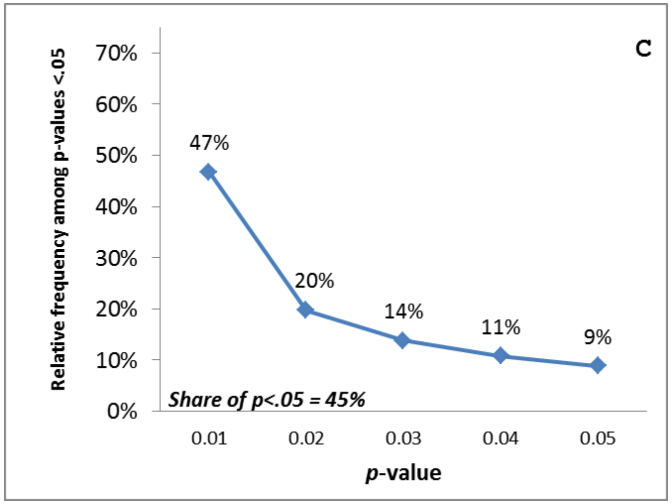


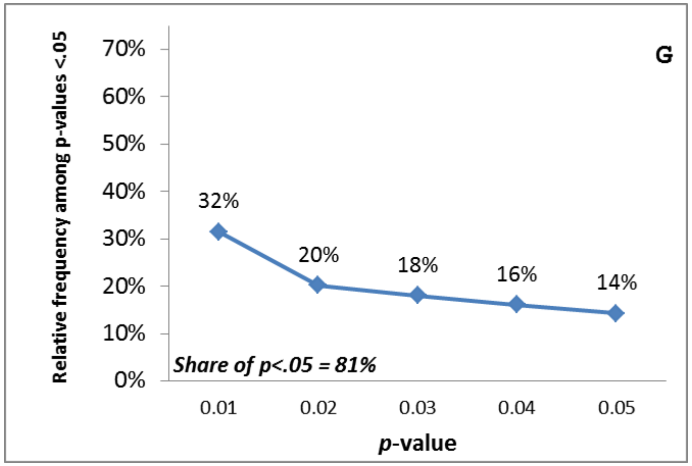
p-hacked findings?

NO | YES

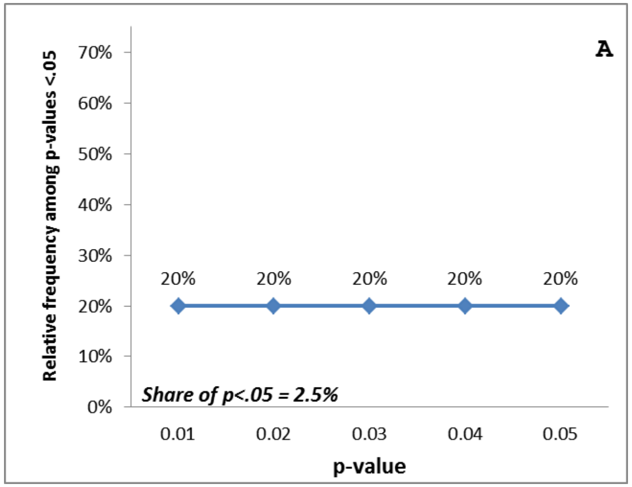
Effect Exists?

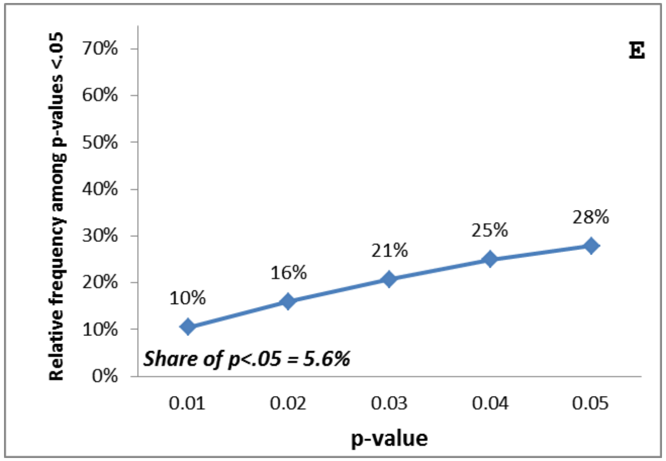
YES





NO





Note:

- *p*-curve does not test if p-hacking happens.

(it "always" does)

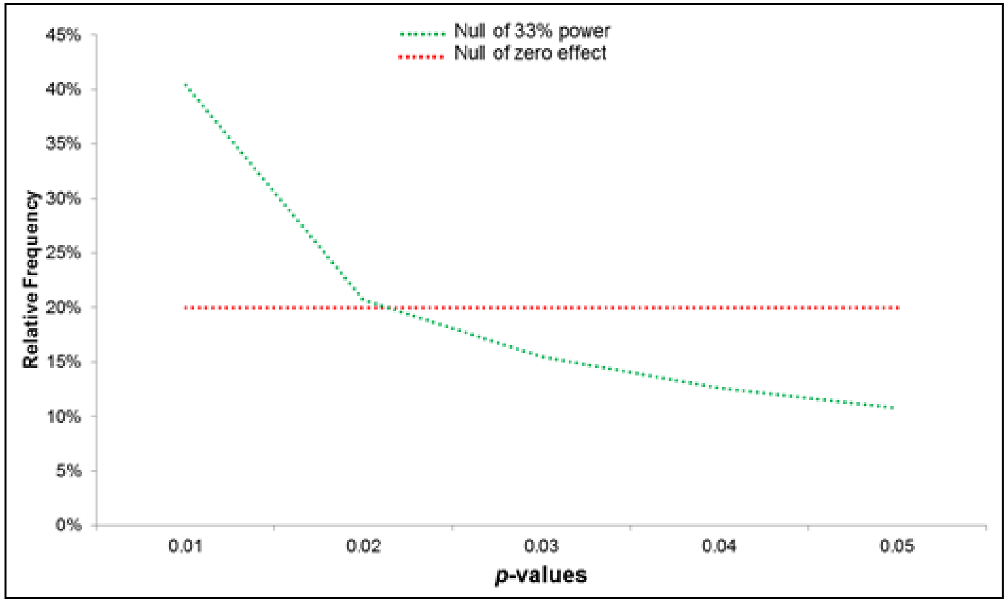
Rather:

- Whether p-hacking was so intense that it eliminated evidential value (if any).

Outline

- Shape
- Inference
- Demonstration
- How often is p-curve wrong?
- Effect-size estimation
- Selecting *p*-values

Inference with p-curve

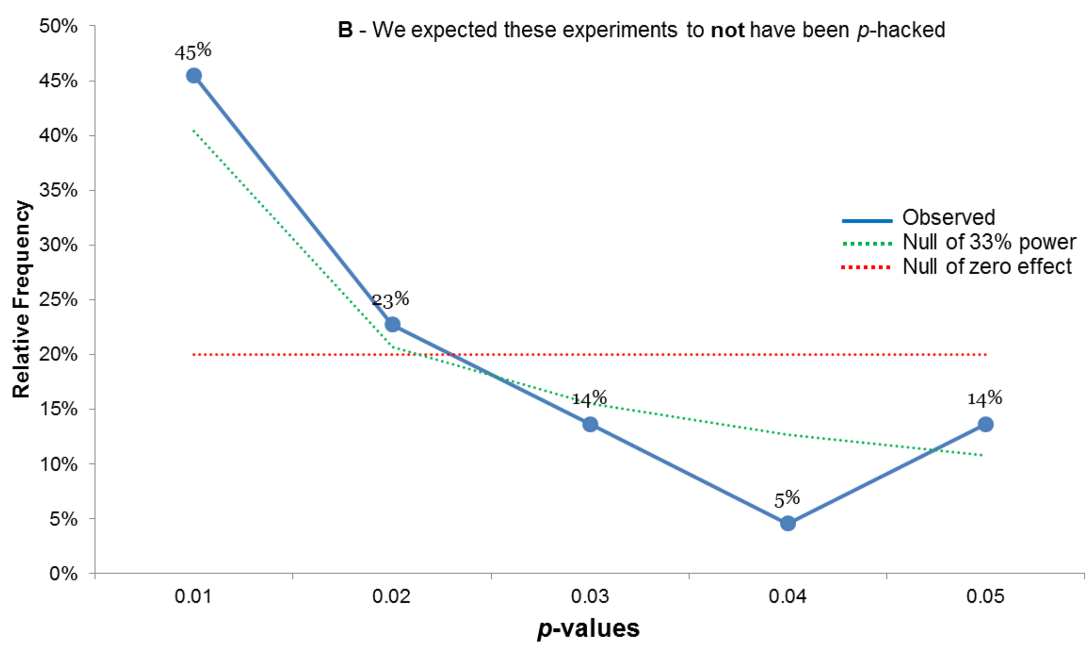


1) Right-skewed?
2) Flatter than studies powered at 33%?
3) Left-skewed?

Outline

- Shape
- Inference
- Demonstration
- How often is p-curve wrong?
- Effect-size estimation
- Selecting *p*-values

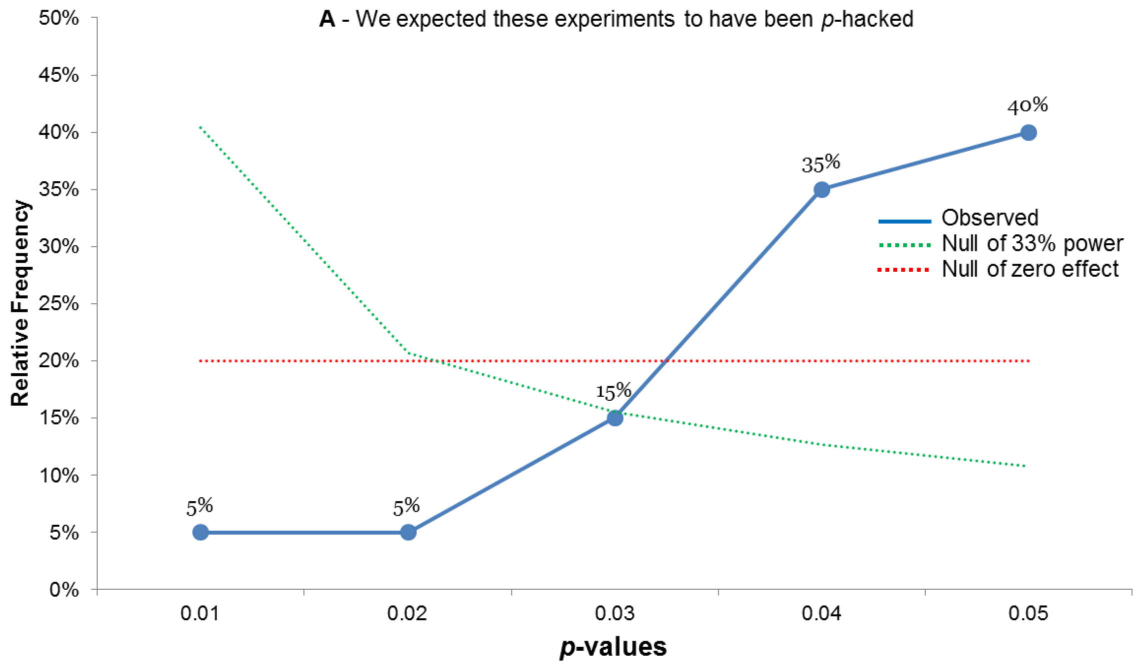
Set 1: JPSP with no exclusions nor transformations



Statistical Inference	*pp*-values	Binominal *p*
1) Studies contain evidential value (right-skewed)	<.0001	.0262
2) Studies lack evidential value (power less than 33%)	.5670	.7124
3) Studies lack evidential value and were intensely *p*-hacked (left-skewed)	.9781	.9915

The observed p-curve includes 22 p-values (16 are p<.025)
An additional 3 were p>.05 and hence excluded.

Set 2: JPSP result reported only with covariate



Statistical Inference	*pp*-values	Binominal *p*
1) Studies contain evidential value (right-skewed)	.9987	.9998
2) Studies lack evidential value (power less than 33%)	.0001	<.0001
3) Studies lack evidential value and were intensely *p*-hacked (left-skewed)	.0311	.0013

The observed p-curve includes 20 p-values (3 are p<.025)
An additional 3 were p>.05 and hence excluded.

- **Next**: New Example

One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects[†]

By Zacharias Maniadis, Fabio Tufano, and John A. List[*]

Some researchers have argued that anchoring in economic valuations casts doubt on the assumption of consistent and stable preferences. We present new evidence that explores the strength of certain anchoring results. We then present a theoretical framework that provides insights into why we should be cautious of initial empirical findings in general. The model importantly highlights that the rate of false positives depends not only on the observed significance level, but also on statistical power, research priors, and the number of scholars exploring the question. Importantly, a few independent replications dramatically increase the chances that the original finding is true. (*JEL* D12, C91)

First draft: 2013 10 24
This draft: 2014 04 11

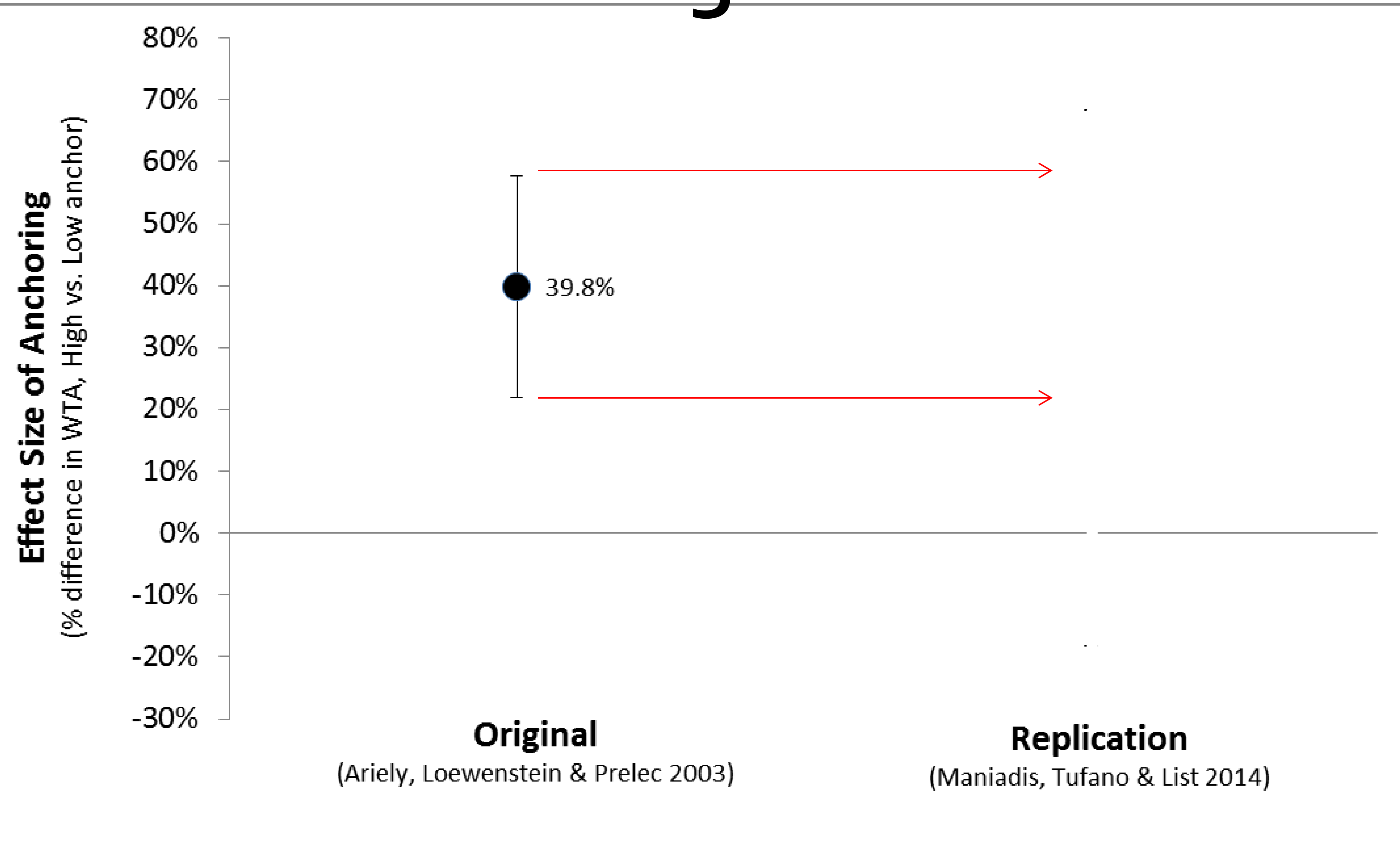
Anchoring is Not a False-Positive: Maniadis, Tufano, and List's (2014) "Failure-to-Replicate" is Actually Entirely Consistent with the Original

Uri Simonsohn
University of Pennsylvania
uws@wharton.upenn.edu

Joseph P. Simmons
University of Pennsylvania
jpsimmo@wharton.upenn.edu

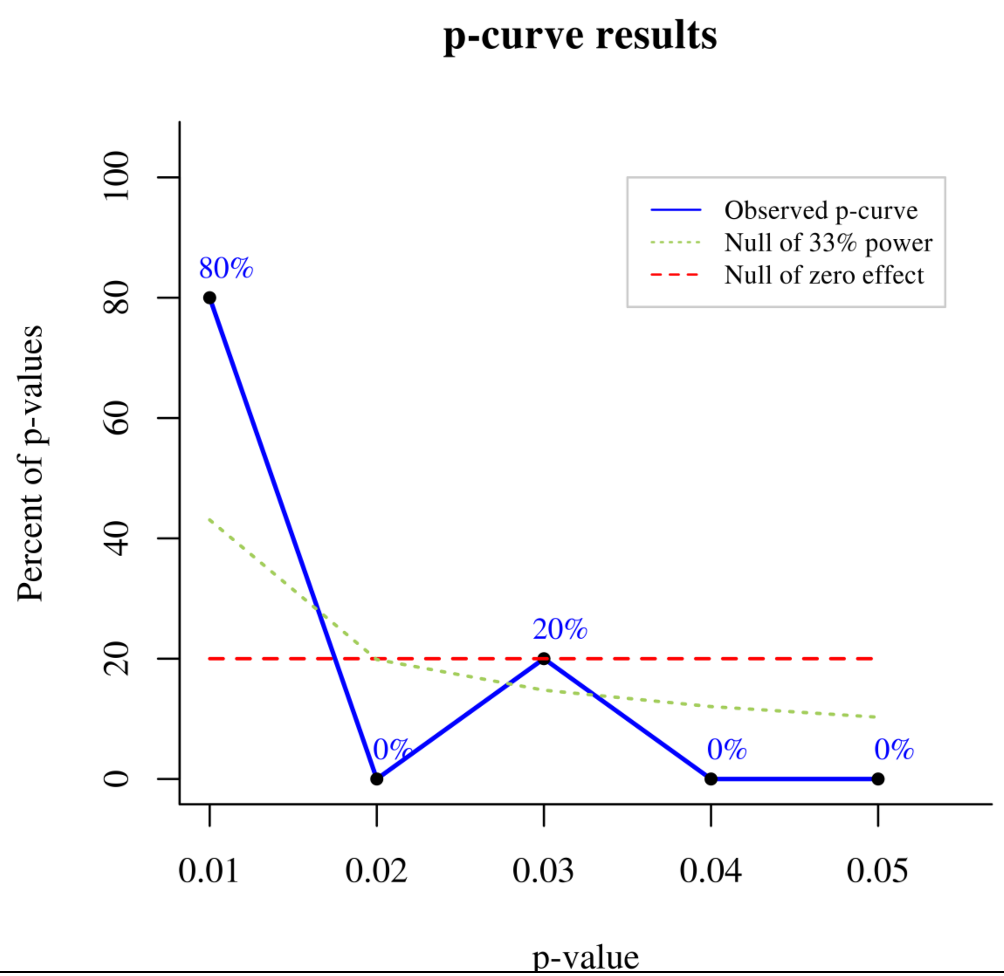
Leif D. Nelson
UC Berkeley
leif_nelson@haas.berkeley.edu

Anchoring and WTA



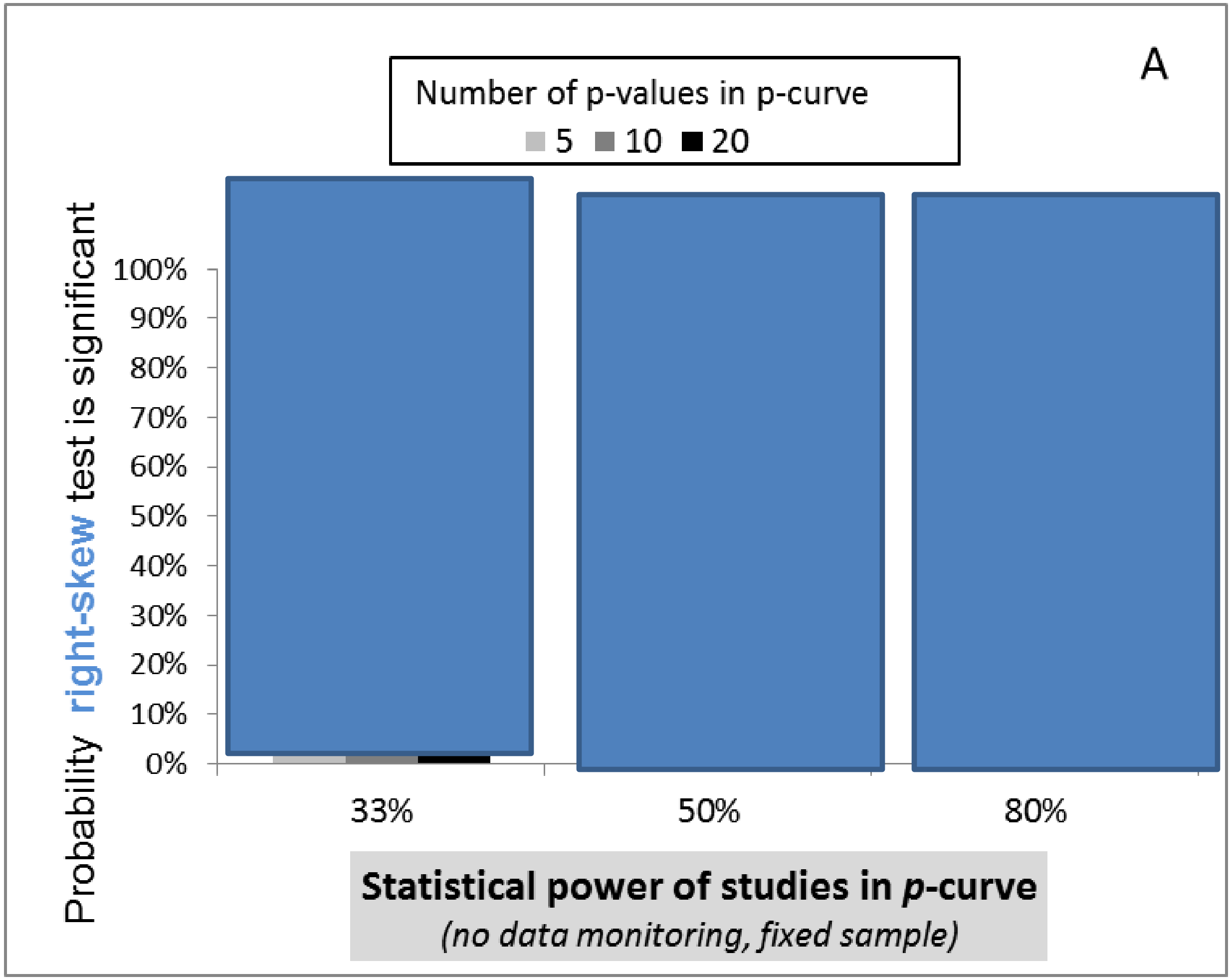
- Bad replication $\neg\rightarrow$ Good original

- Was original a false-positive?



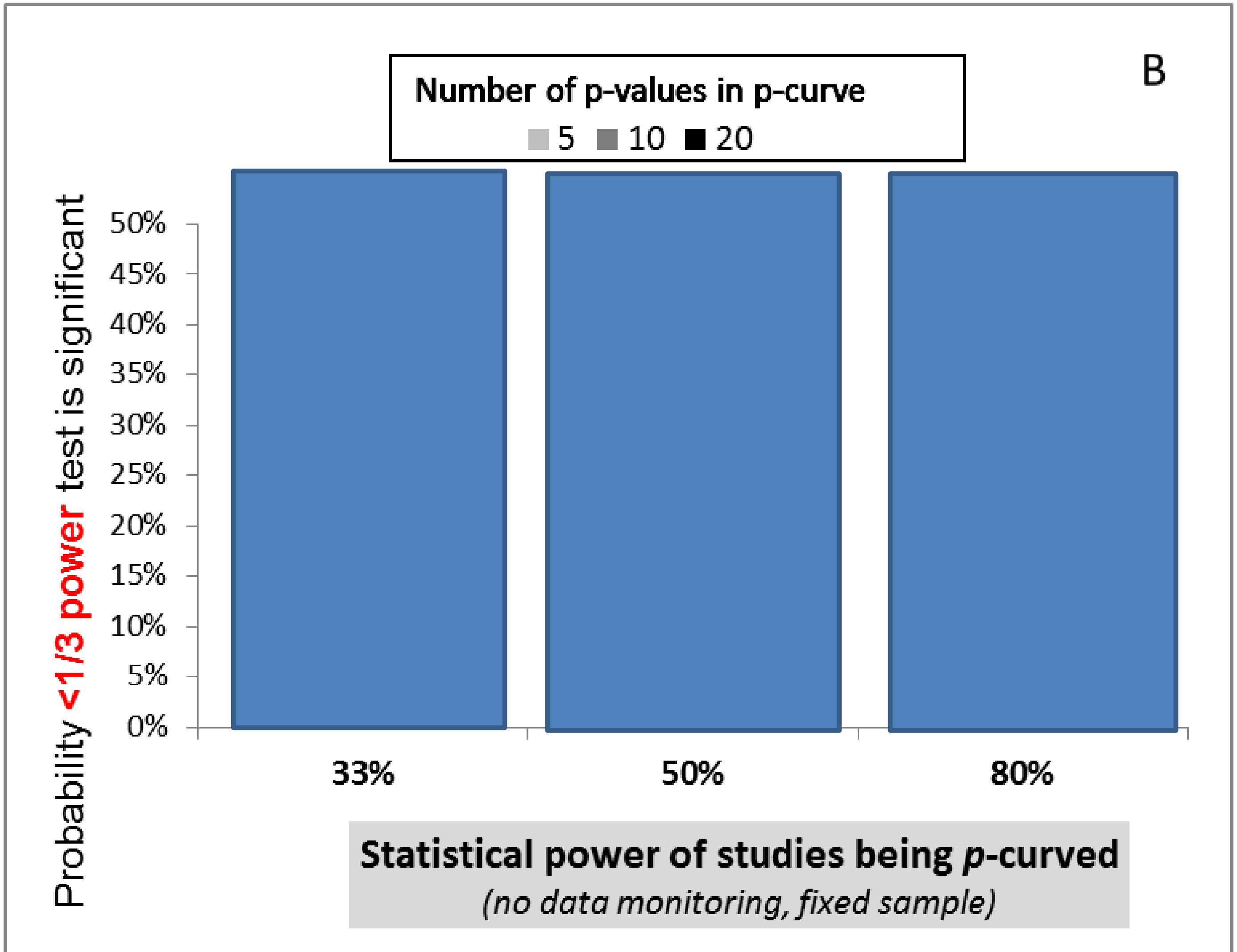
Statistical Inference	Results
1) Studies contain evidential value *(right-skewed)*	$\chi^2(10)=33.76$, $p=.0002$
2) Studies lack evidential value *(flatter than 33% power)*	$\chi^2(10)=2.8$, $p=.9857$
3) Studies lack evidential value and were intensely p-hacked *(left-skewed)*	$\chi^2(10)=1.35$, $p=.9993$

When effect exists, how often does *p*-curve say "evidential value"



Highlights:
More power at 5
Certain with 80%

When effect exists, how often does *p*-curve say "no evidential value"



Highlights

- P-curve is 'never' wrong on properly powered studies.

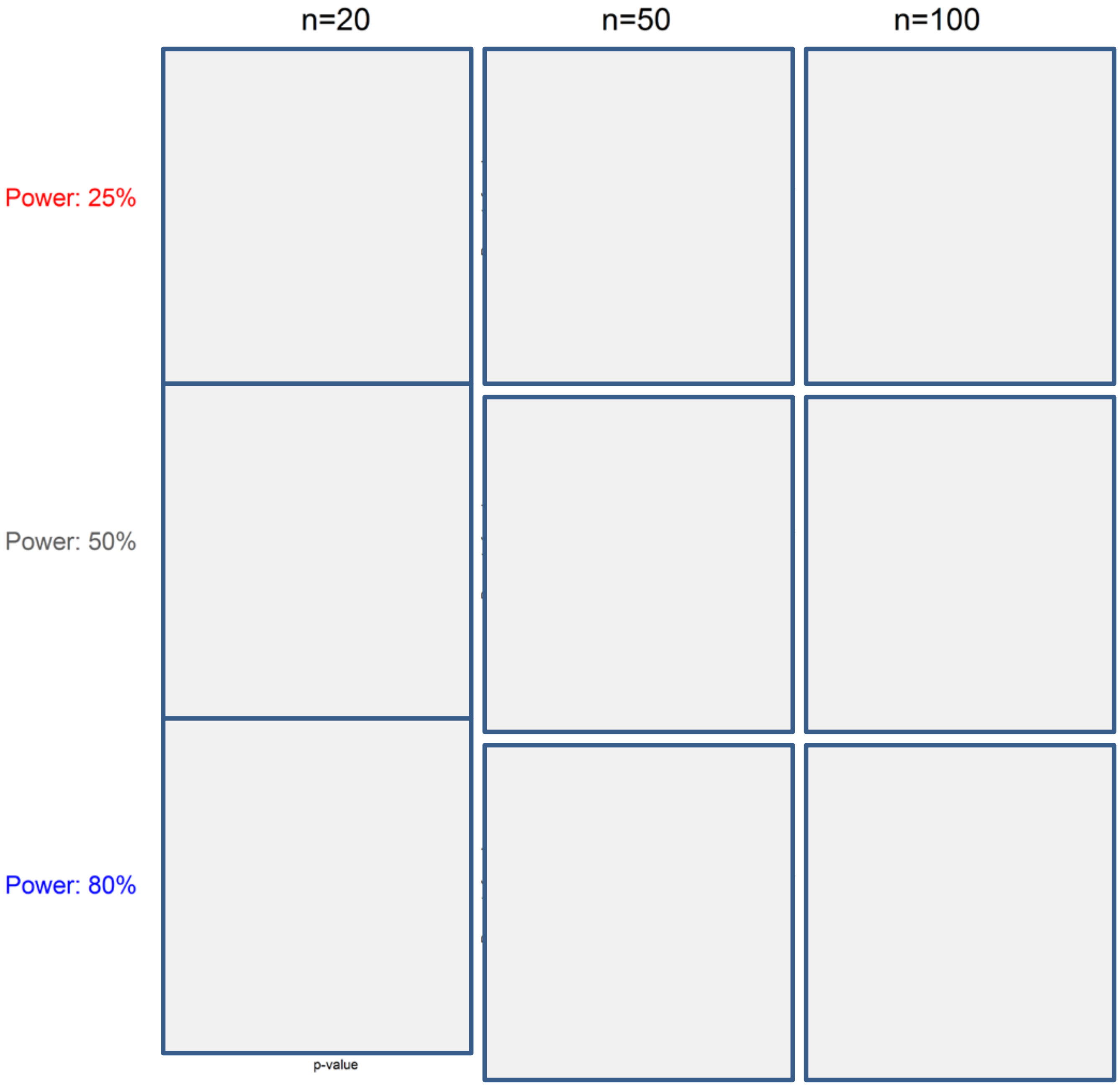
Broad big picture applications

- Possible uses:
  - Meta-analyses of X on Y
  - Meta-analyses of X on anything
  - Meta-analyses of anything on Y
  - Relative truth of opposing findings
    - X is good for Y, vs
    - X is bad for Y
  - Is this journal, on average, true?
  - Universities vs. pharmaceuticals

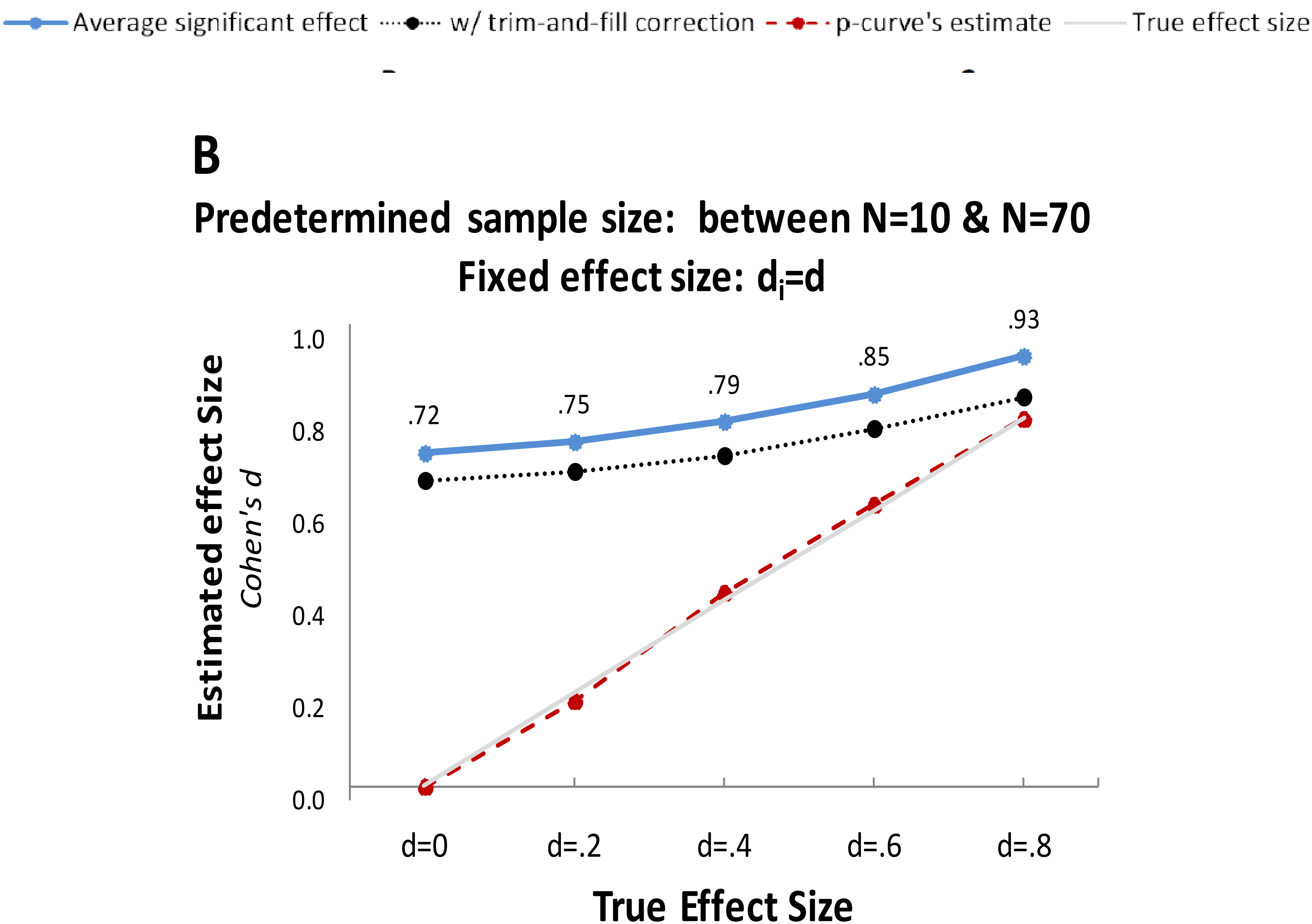
Everyday applications

(note: 5 p-values can be plenty)

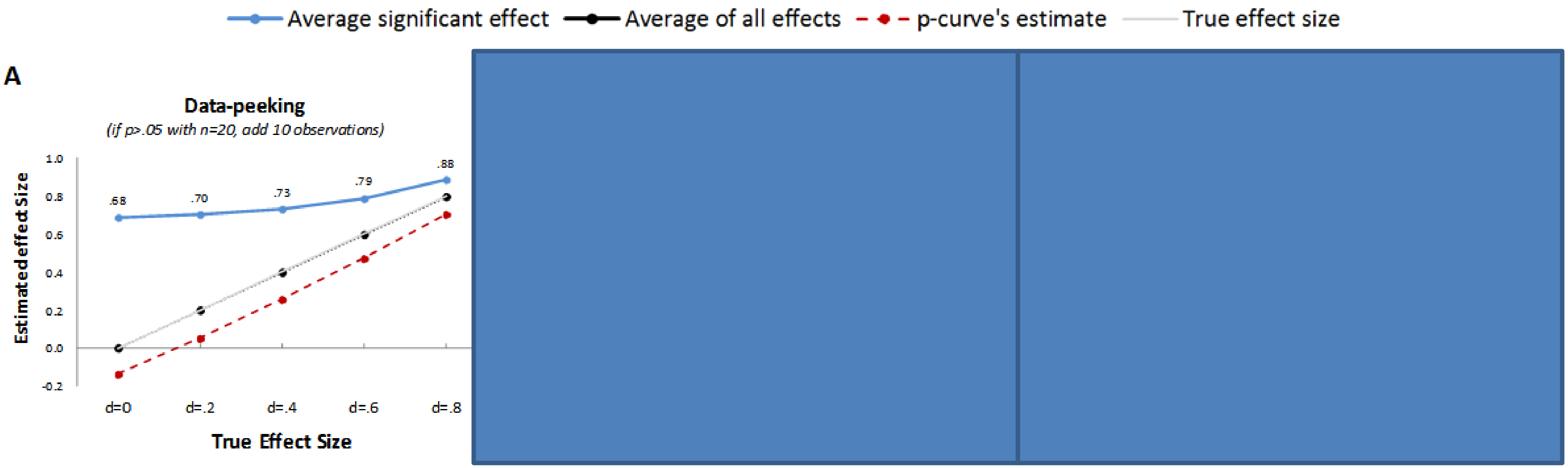
- **Reader**: Should I read this paper?
- **Researcher**: Run expensive follow-up?
- **Researcher**: Explain inconsistent previous finding
- **Reviewer**: Ask for direct replications?


n=20
n=50
n=100
Power: 25%
Power: 50%
Power: 80%
p-value

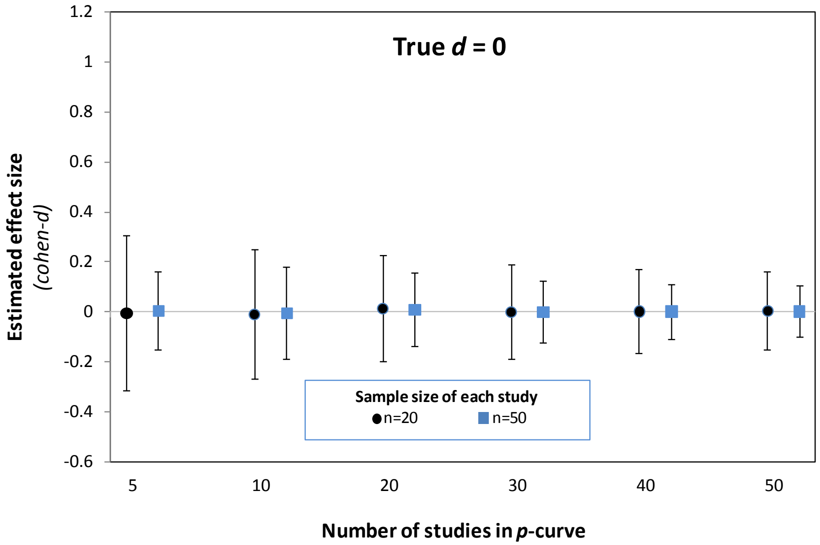
- **Next**.
  - Simulated meta-analysis, file-drawering studies.


Average significant effect
w/ trim-and-fill correction
p-curve's estimate
True effect size
B
Predetermined sample size: between N=10 & N=70
Fixed effect size: d_i=d
Estimated effect Size
Cohen's d
1.0
0.8
0.6
0.4
0.2
0.0
.72
.75
.79
.85
.93
d=0
d=.2
d=.4
d=.6
d=.8
True Effect Size

- **Next**.
  - Simulated meta-analysis, *p*-hacking


Average significant effect
Average of all effects
p-curve's estimate
True effect size
A
Data-peeking
(if p>.05 with n=20, add 10 observations)
Estimated effect Size
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
.68
.70
.73
.79
.88
d=0
d=.2
d=.4
d=.6
d=.8
True Effect Size

- **Next**. Precision from few studies



Number of studies in *p*-curve

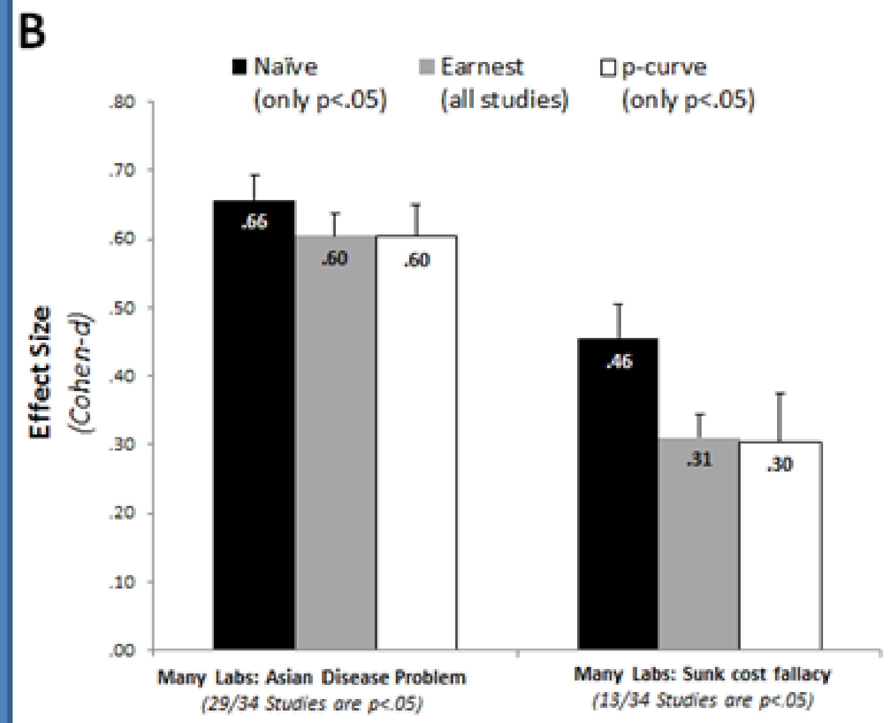
- **Next**. Demonstration 1: Many Labs Replication project
  - Real study, participants, data
  - But, see all attempts

Open Science Framework BETA   Explore   Help   Search

Investigating Variation in Replicability: A “Many Labs” Replication Project

Contributors: Richard A. Klein | Kate Ratliff | Michelangelo Vianello | Reginald B. Adams, Jr. | Stepan Bahnik | Michael
Mark Brandt | Beach Brooks | Claudia Brumbaugh | Zeynep Cemalcilar | Jesse J. Chandler | Winnee Cheong | William
Matthew Eisner | Natalia Frankowska | David Furrow | Elisa Maria Galliani | Fred Hasselman | Joshua A. Hicks | Jame
Jeffrey R. Huntsinger | Hans IJzerman | Melissa-Sue John | Jennifer Joy-Gaba | Heather Kappes | Lacy Elise Krueger
Robyn Mallett | Wendy Morris | Anthony J. Nelson | Jason A. Nier | Grant Packard | Ronaldo Pilati | Abraham M. Rutch
Skorinko | Robert W. Smith | Troy G. Steiner | Justin Storbeck | Lyn van swol | Donna Thompson | Anna van 't Veer | L
Aaron Wichman | Julie A. Woodzicka | Brian A. Nosek

- 36 labs
- 13 "effects"
  - Example 1. Sunk Cost (Significant: 50% labs)
  - Example 2. Asian Disease (86%)

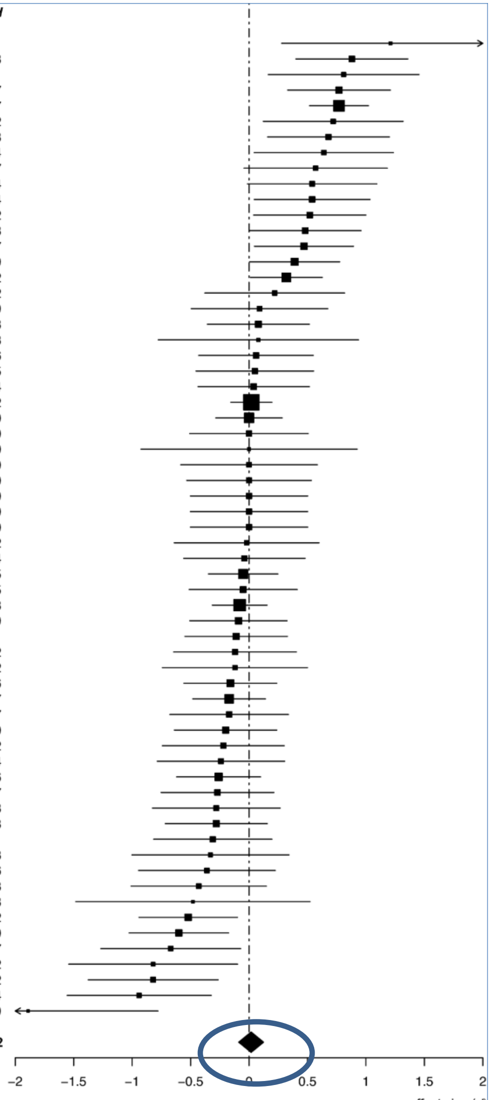

B
Naïve (only p<.05)
Earnest (all studies)
p-curve (only p<.05)
Effect Size (Cohen-d)
.80
.70
.60
.50
.40
.30
.20
.10
.00
.66
.60
.60
.46
.31
.30
Many Labs: Asian Disease Problem
(29/34 Studies are p<.05)
Many Labs: Sunk cost fallacy
(13/34 Studies are p<.05)

- **Next**. Demonstration 2: Choice Overload

A demonstration

Choice Overload meta-analysis

Author	Detail	Year	*N*	weight	*d*
Inbar, Hanko, Botti, & Gilovich	study 1, time pressure	2008	21	0.4%	1.21
Iyengar & Lepper	chocolate study	2000	67	1.4%	0.88
Greifeneder, Scheibehenne, & Kleber	experiment 1, 6 attributes	2010	40	0.8%	0.81
Shah & Wolford	–	2007	80	1.6%	0.77
Iyengar & Lepper	jam study	2000	249	4.9%	0.77
Chernev	study 1, no ideal point	2003	43	0.9%	0.72
Reutskaja & Hogarth	–	2009	60	1.2%	0.68
Mogilner, Rudnick, & Iyengar	preference constructors	2008	60	0.9%	0.64
Chernev	study 2, no ideal point	2003	41	0.8%	0.57
Greifeneder, Scheibehenne, & Kleber	experiment 2, 9 attributes	2010	52	1.0%	0.54
Diehl & Poynor	study 2	2007	65	1.3%	0.54
Gingras	study 4	2003	69	1.4%	0.52
Haynes	–	2009	69	1.4%	0.48
Chernev	study 3, low ideal point score	2003	86	1.7%	0.47
Scheibehenne, Greifeneder, & Todd	charity study 3	2009	115	2.1%	0.39
Diehl & Poynor	study 3	2007	165	3.3%	0.32
Gao & Simonson	study 4, select–buy	2008	43	0.9%	0.22
Mogilner, Rudnick, & Iyengar	preference matchers	2008	60	0.9%	0.09
Lin & Wu	–	2006	82	1.6%	0.08
Inbar, Hanko, Botti, & Gilovich	study 1, no time pressure	2008	21	0.4%	0.08
Fasolo, Carmeci, & Misuraca	study 1	2009	64	1.3%	0.06
Kleinschmidt	unconstrained set	2008	61	1.2%	0.05
Lenton & Stewart	–	2007	68	1.4%	0.04
Scheibehenne	jam study	2008	504	10.2%	0.02
Scheibehenne & Todd	–	2009	191	3.9%	0.00
Reutskaja	study 2	2008	60	1.2%	0.00
Kahn & Wansink	study 1, disorganized	2004	18	0.4%	0.00
Kahn & Wansink	study 2, disorganized	2004	45	0.9%	0.00
Kahn & Wansink	study 5, disorganized	2004	54	1.1%	0.00
Gingras	study 2, expertise	2003	61	1.2%	0.00
Gingras	study 2, no expertise	2003	61	1.2%	0.00
Gingras	study 3	2003	61	1.2%	0.00
Kilchherr, Wänke & Messner	sundae study, simultaneous	2008	40	0.8%	−0.02
Scheibehenne, Greifeneder, & Todd	charity study 1, 5 vs. 40	2009	57	1.2%	−0.04
Scheibehenne, Greifeneder, & Todd	music study, US sample	2009	174	3.5%	−0.05
Haynes & Olson	study 2	2007	72	1.5%	−0.05
Scheibehenne	wine study	2008	280	5.7%	−0.08
Lenton, Fasolo, & Todd	–	2008	89	1.8%	−0.09
Scheibehenne, Greifeneder, & Todd	restaurant study	2009	80	1.6%	−0.11
Inbar, Hanko, Botti, & Gilovich	study 2	2008	56	1.1%	−0.12
Greifeneder, Scheibehenne, & Kleber	experiment 1, 1 attribute	2010	40	0.8%	−0.12
Scheibehenne, Greifeneder, & Todd	charity study 2, USA	2009	112	2.0%	−0.16
Scheibehenne, Greifeneder, & Todd	music study, German sample	2009	160	3.2%	−0.17
Kleinschmidt	constrained set	2008	60	1.2%	−0.17
Greifeneder	–	2008	80	1.6%	−0.20
Söllner & Newell	perfume study	2009	57	1.2%	−0.22
Alleman, et al.	–	2007	52	1.0%	−0.24
Fasolo, Carmeci, & Misuraca	study 2	2009	120	2.4%	−0.26
Scheibehenne	jelly bean study	2008	66	1.3%	−0.27
Greifeneder, Scheibehenne, & Kleber	experiment 2, 4 attributes	2010	52	1.0%	−0.28
Chernev	study 3, high ideal point score	2003	81	1.6%	−0.28
Scheibehenne, Greifeneder, & Todd	charity study 1, 2 vs. 30	2009	60	1.2%	−0.31
Chernev	study 2, ideal point available	2003	34	0.7%	−0.33
Chernev	study 1, ideal point available	2003	45	0.9%	−0.36
Olivola & Schwartz	–	2006	46	0.9%	−0.43
Effron & Lepper	–	2007	16	0.3%	−0.48
Berger, Draganska, & Simonson	study 3	2007	90	1.8%	−0.52
Gingras	study 1	2003	89	1.7%	−0.60
Gao & Simonson	study 4, buy–select	2008	43	0.9%	−0.67
Söllner & Newell	sun cream study	2009	32	0.6%	−0.82
Kahn & Wansink	study 5, organized	2004	54	1.0%	−0.82
Kahn & Wansink	study 2, organized	2004	45	0.8%	−0.94
Kahn & Wansink	study 1, organized	2004	18	0.3%	−1.89
Summary			***N* = 5036**		***D* = 0.02**

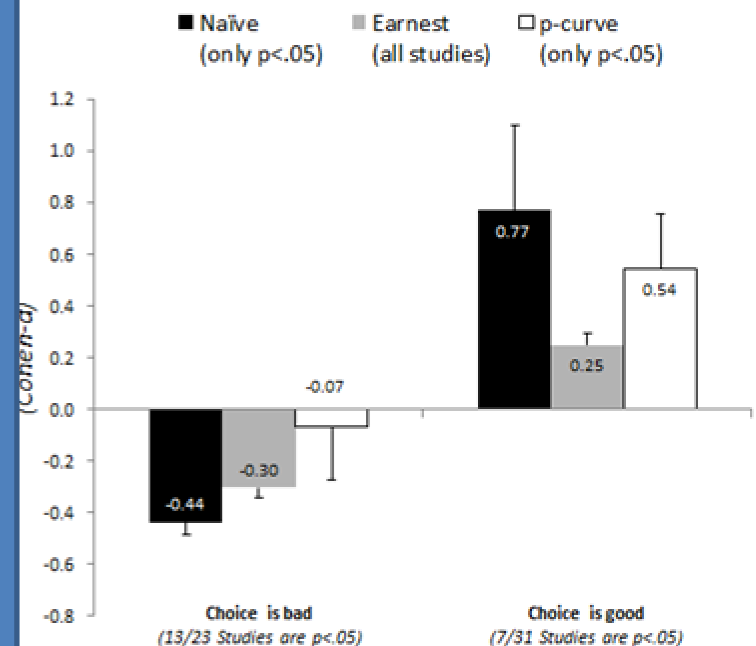


**

Choice is bad

Choice is good

**


Naïve (only p<.05)
Earnest (all studies)
p-curve (only p<.05)
(Cohen-d)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
0.77
0.25
0.54
-0.44
-0.30
-0.07
Choice is bad
(13/23 Studies are p<.05)
Choice is good
(7/31 Studies are p<.05)

How to think about p-values

- When a study has lots of statistical power (big effect + big sample), expect to see very small p-values.
- When you see a really big p-value (p = .048), you should be concerned.
- Unexpected thought: When the p-values are really small in the absence of statistical power, you can have different (more unsettling) concerns.

I don't have any more slides, but I have many more thoughts and opinions.

Ask.

datacolada.org

P-curve

p-curve.com

Paper 1 - Evidential Value	Paper 2 - Effect size	The online app	The User Guide	Supp Materials