Statistical Methods for Replicability Assessment

Kenneth Hung and Will Fithian

University of California, Berkeley

{kenhung, wfithian}@berkeley.edu

March 12, 2019

Table of Contents

Introduction: Replicability Crisis

Many Psychology Findings Not as Strong as Claimed, Study Says

By Benedict Carey

Aug. 27, 2015

The past several years have been bruising ones for the credibility of the social sciences. A star social psychologist was caught fabricating data, leading to more than 50 retracted papers. A top journal published a study supporting the existence of ESP that was widely criticized. The journal Science pulled a political science paper on the effect of gay canvassers on voters' behavior because of concerns about faked data.

Now, a painstaking yearslong effort to reproduce 100 studies published in three leading psychology journals has found that more than half of the findings did not hold up when retested. The analysis was done by

Introduction: Replicability crisis

Social psychology facing urgent crisis in replicability of results

Commonly attributed to varied factors

- selection for significance
- *p*-hacking, questionable research practices (QRPs)
- fraud
- infidelity of replication experimental designs
- flaws in original experimental designs
- "Hidden moderators": subtle, uncontrollable differences in experimental conditions

Reproducibility Project: Psychology

- Preregistered replications of 100 studies published in 2008 in three top psych. journals
- Massive collaborative effort by hundreds of researchers

Results from RP:P

RP:P reported descriptive statistics:

- 36% of replications significant in same direction as original study
- 47% of original point estimates in replication studies' 95% CIs
- 83% of the effect size estimates declined ($\hat{\theta}_{i,R}/\hat{\theta}_{i,O} < 1$)

Widely reported as damning result:

- *Washington Post*: "... affirms that the skepticism [of published results] was warranted" (Achenbach, 2015)
- *Economist*: "... managed to replicate satisfactorily the results of only 39% of the studies investigated" (Ano, 2016)
- *New York Times*: "more than half of the findings did not hold up when retested" (Carey, 2015)

Debate

What should we make of these numbers?

- e.g. 47% of original point estimates in replication studies' 95% CIs

Gilbert et al. (2016a) critiqued 47% number:

- Confidence interval $\neq$ predictive interval
- No replication is exact ($\theta_{i,O} \neq \theta_{i,R}$)
- Low fidelity of some replications (e.g. race questionnaire in Italy)
- "OSC seriously underestimated the reproducibility of psychological science"

Further debate between defenders (Anderson et al., 2016; Srivastava, 2016; Nosek and Gilbert, 2016), critics (Gilbert et al., 2016c,b)

What estimand?

Did OSC "underestimate replicability?"

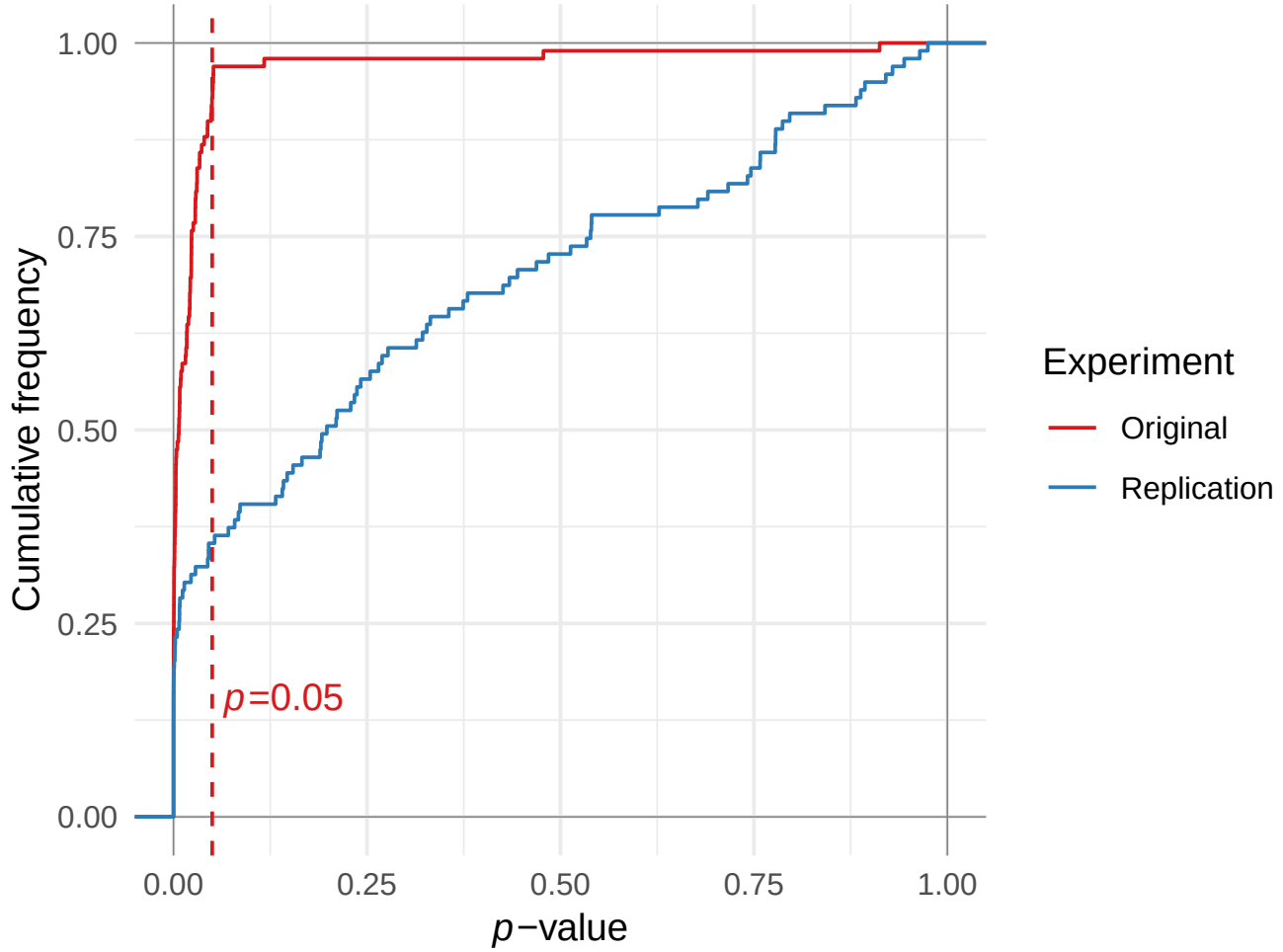
First need to answer: what is the estimand?

RP:P offers

- descriptive statistics
- tiny p-values for tests of very strong nulls
  - e.g. McNemar's test of whether orig. studies more likely to be significant at level 0.05
- no attempt to define target of inference
- no attempt to disentangle sources of error

Selection for significance

RP:P data: unmistakable sign of selection at $\alpha = 0.05$



Can this alone explain all results?

Simulation: Selection bias

Can selection bias alone explain RP:P's descriptive statistics?

Simulation experiment

- all original / replication studies: same effect size θ
- Gaussian estimators $\hat{\theta}_{i,O}, \hat{\theta}_{i,R}$, s.e. $= 1$
- observe pair only when $|\hat{\theta}_{i,O}| > z_{\alpha/2}$

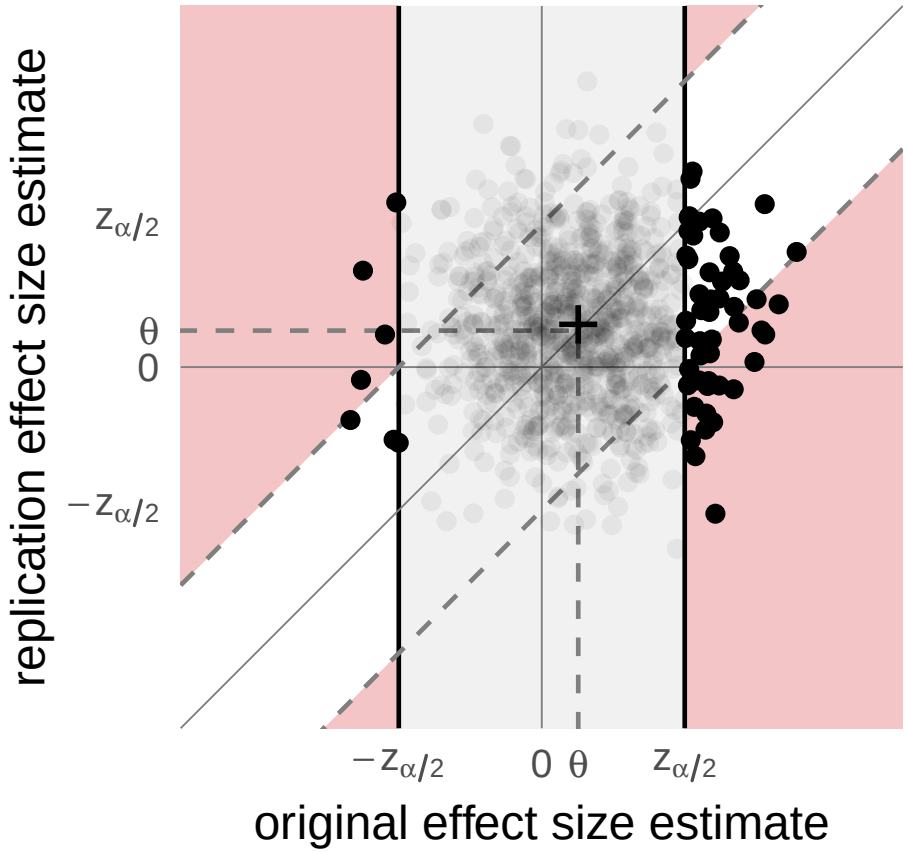
Plot descriptive statistics as function of θ

Simulation: Selection bias

Question: with $\theta = 1/2$, what fraction of repl. CIs cover orig. estimates?

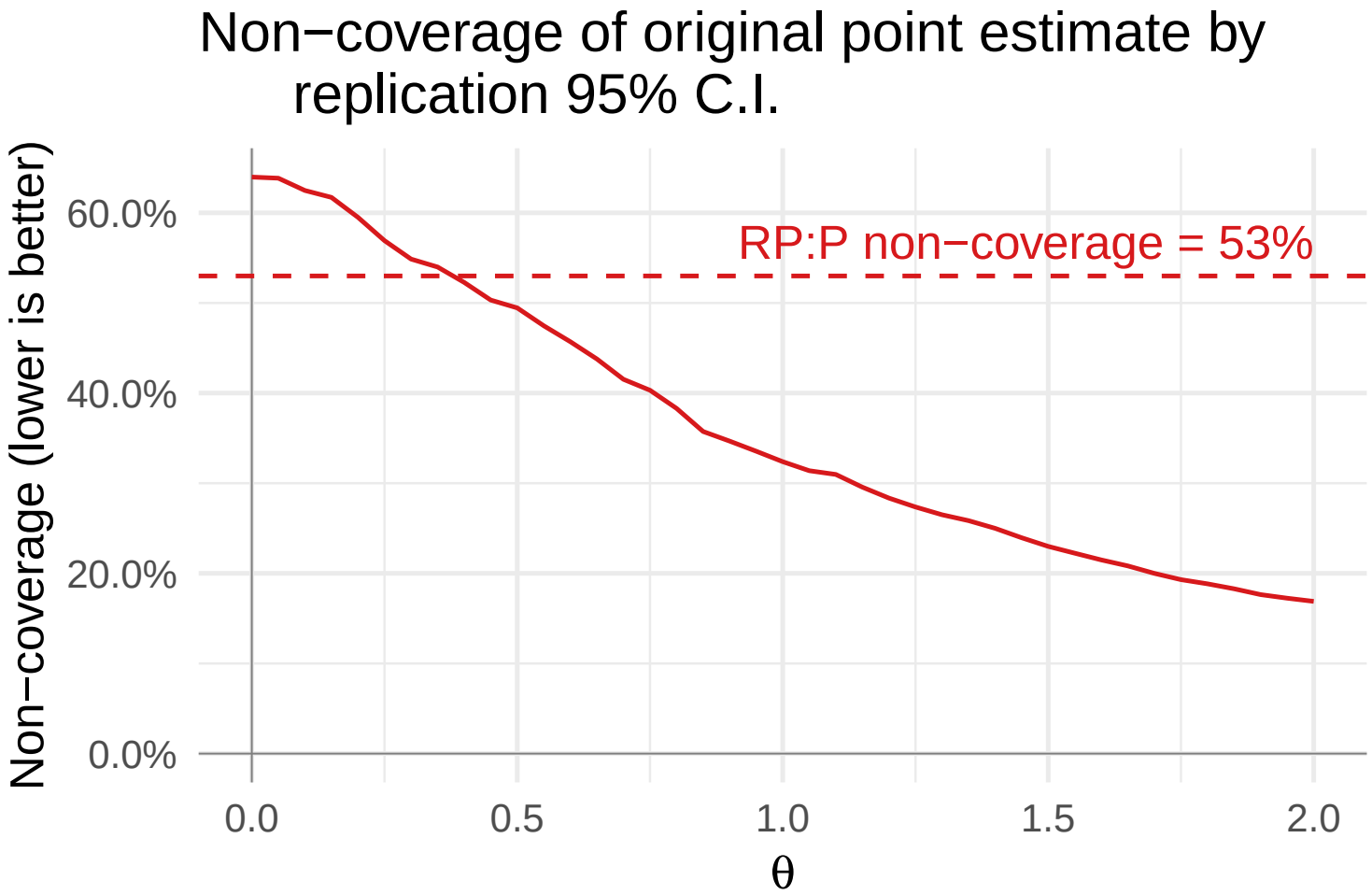
Simulation: Selection bias

Question: with $\theta = 1/2$, what fraction of repl. CIs cover orig. estimates?

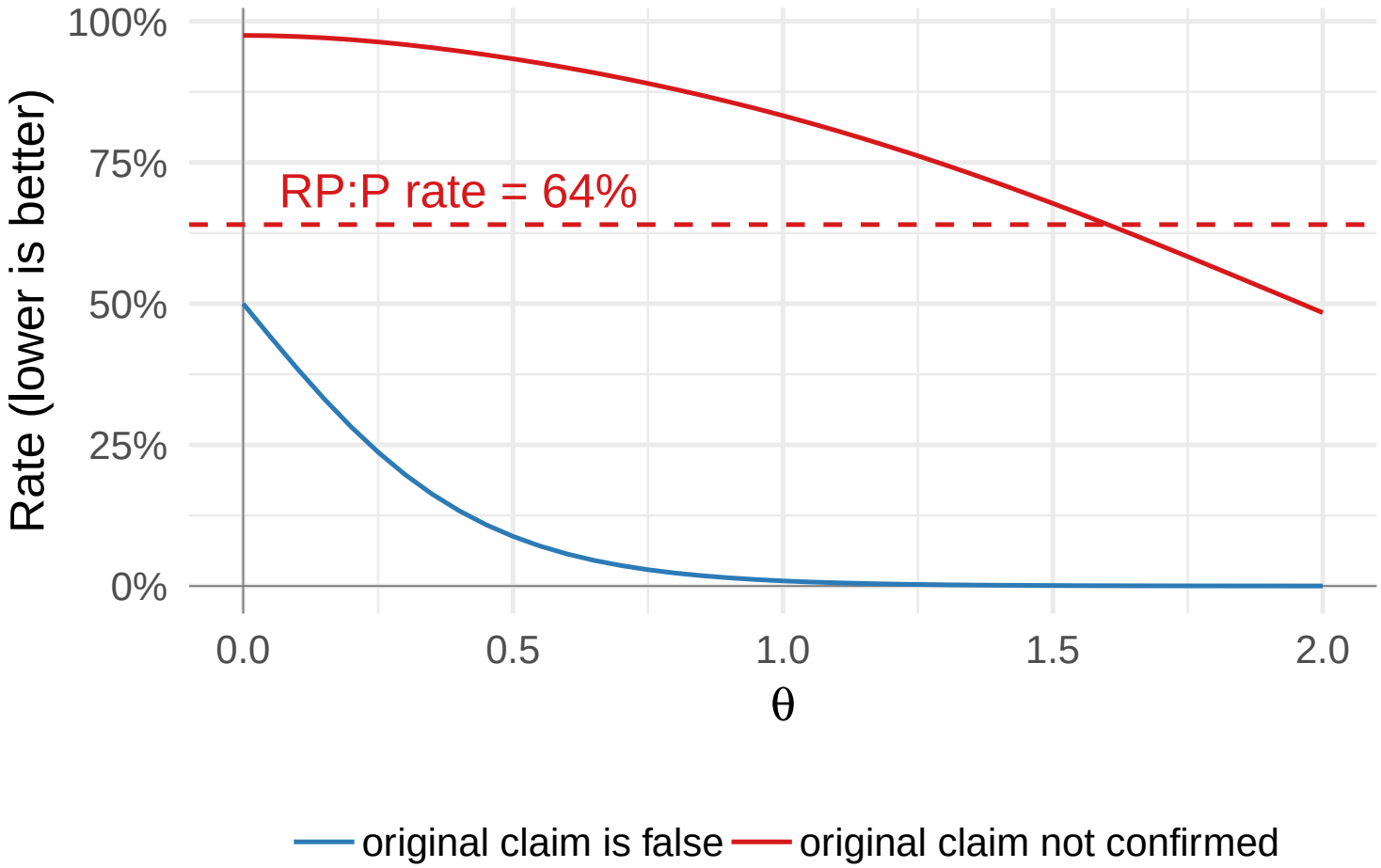


Answer: $\approx 50\%$

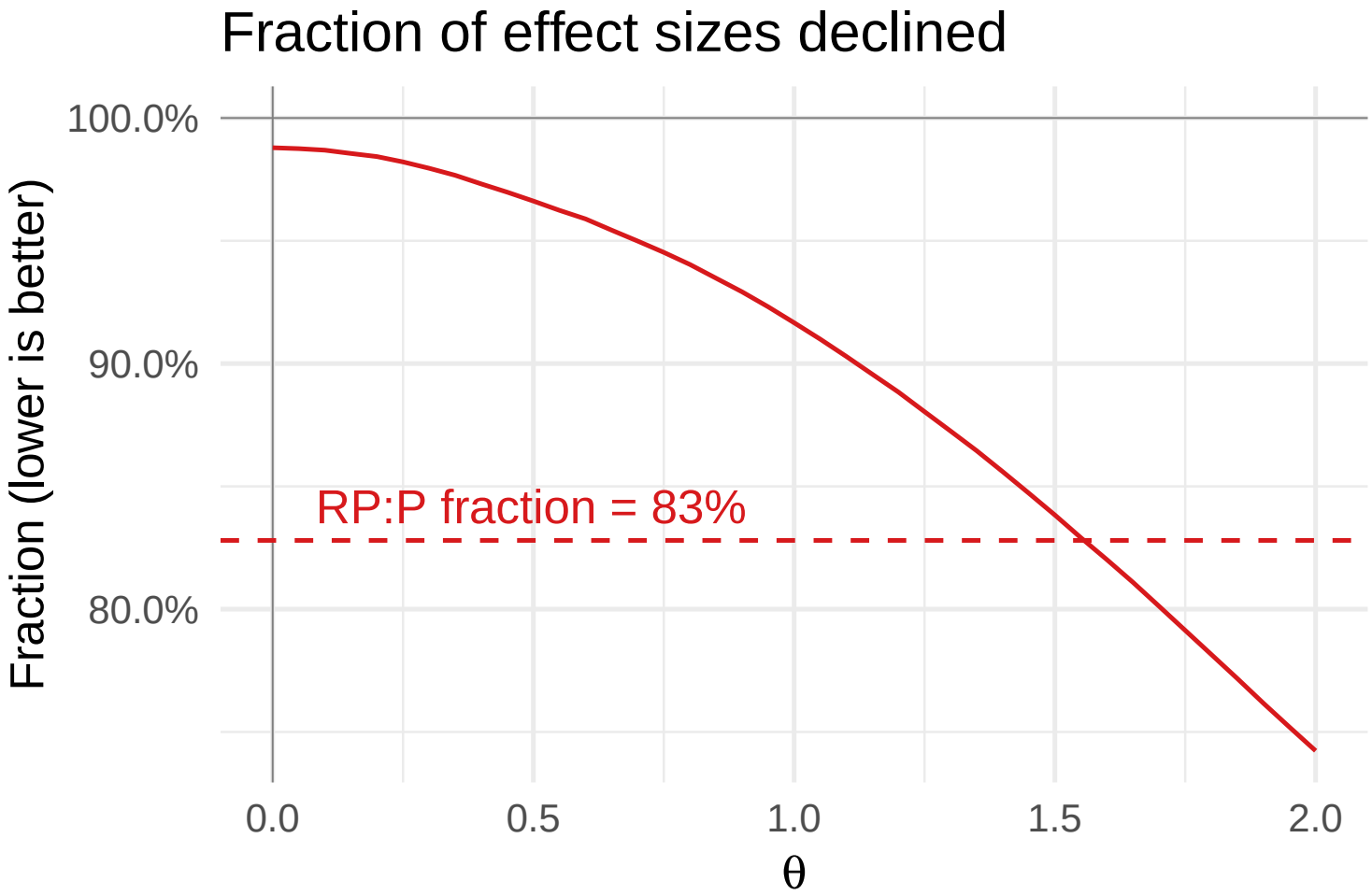
Simulation: Selection bias



Simulation: Selection bias



Simulation: Selection bias



Selection bias

Selection bias is basic feature of data

- Qualitatively, can explain RP:P metrics
- Can't learn anything else from data unless we adjust for it

Different explanations suggest different priorities for reform, e.g.:

- Selection bias: publish negative results, post-hoc stat. adjustment
- Infidelities: more detailed methods sections?
- Hidden moderators: abandon experimental psychology?

Key tools:

- conditional post-selection inference (Lee et al., 2016; Fithian et al., 2014, many others)
- ideas from multiple testing (Benjamini and Hochberg, 1995; Storey, 2002; Heller et al., 2007, many others)

Table of Contents

A model for replications

Model for original (O) / replication (R) study pair $i = 1, \ldots, m$:

$$\hat{\theta}_{i,O} \sim N\left(\theta_{i,O},\ \sigma^2_{i,O}\right) 1_{\{|\hat{\theta}_{i,O}|>c\}} \quad \text{and} \quad \hat{\theta}_{i,R} \sim N\left(\theta_{i,R},\ \sigma^2_{i,R}\right) \tag{1}$$

Can formalize three definitions in terms of parameters of model (1)

- Some hypotheses defined in terms of $S_i = \text{sign}(\hat{\theta}_{i,O})$

Finite population model:

- no assumptions on dist. of $(\theta_{i,O}, \theta_{i,R})$
- agnostic to why $\theta_{i,O} \neq \theta_{i,R}$

Defining replicability

What do we estimate when we "estimate replicability?"

RP:P statistic 1: 36% of replications significant in same direction as original study

Definition 1: False directional claims

- *What fraction of original directional claims were wrong?*
- Psychology as enterprise in large-scale multiple testing
- *Type S error:* true effect 0 or opposite sign as claimed (Gelman and Tuerlinckx, 2000)
- Would it be different if we used a lower publication threshold?

Answers question: "Would an exact replication with huge n affirm the directional claim?"

Formalizing replicability

Definition 1: False directional claims

What fraction of original directional claims were wrong?

Null hypothesis for Type S error:

$$H_i^{S,O}: \quad S_i \cdot \theta_{i,O} \leq 0$$

Directional FDP for all experiments with $p_{i,O} < \alpha$:

$$\text{FDP}_\alpha = \frac{\#\{i : H_i^{S,O} \text{ true}, \ p_{i,O} < \alpha\}}{\#\{i : p_{i,O} < \alpha\}}$$

Does it improve if $\alpha = 0.005$ were used instead? (Benjamin et al., 2018)

Defining replicability

What do we estimate when we "estimate replicability?"

RP:P statistic 2: 47% of orig. point estimates in repl. studies' 95% CIs

Definition 2: Effect shift of replication

- *How much do effect sizes shift from original to replication?*
- Stability across direct replications
- Bare minimum form of external validity
- Identify studies where effect definitely shifted, produce CIs for shifts

Answers question: "Can psychologists successfully replicate experimental conditions?"

Formalizing replicability

Definition 2: Effect shift of replication

How much do effect sizes shift from original to replication?

Construct CIs for $\theta_{i,O} - \theta_{i,R}$

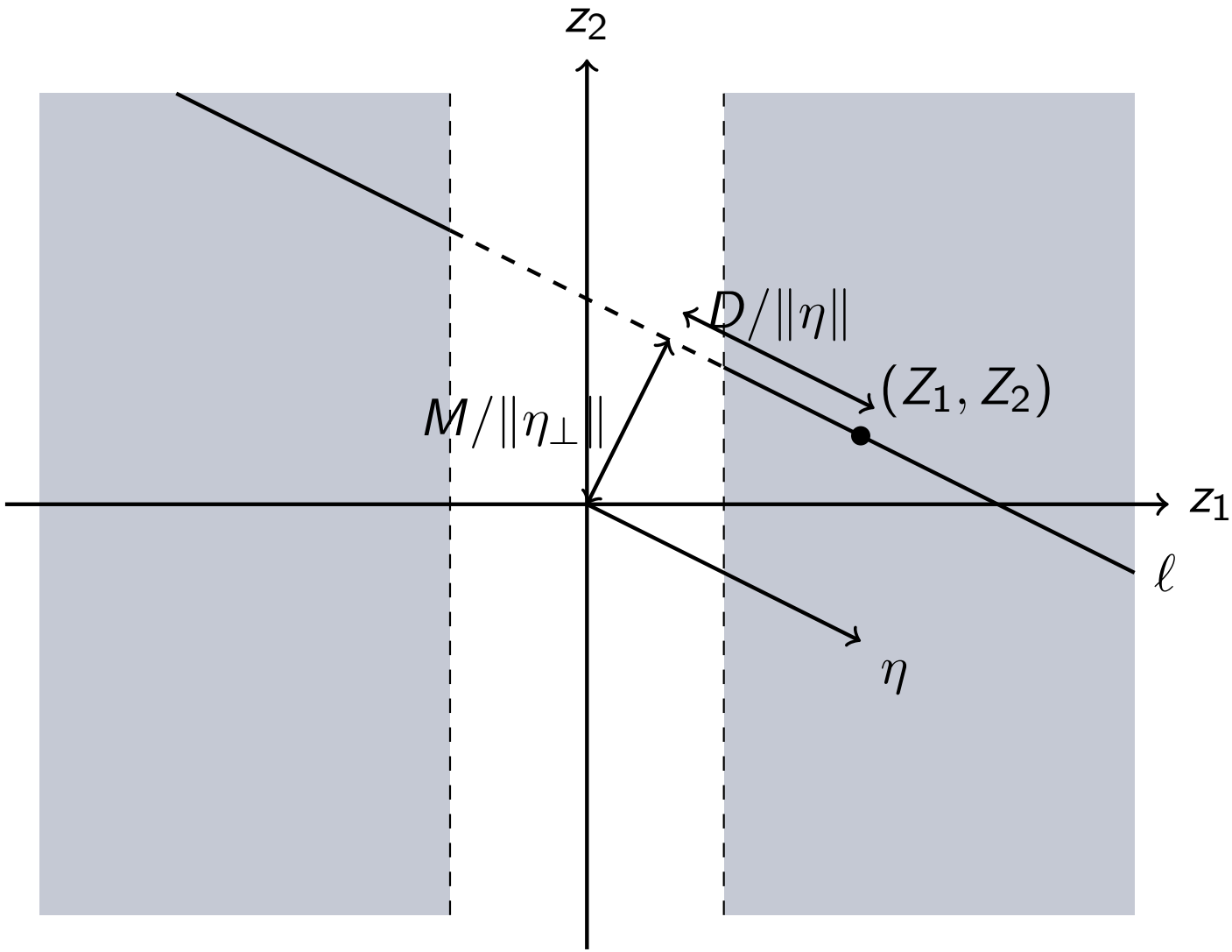
Linear in natural parameters for truncated bivariate Gaussian family

Invert selective z-test (Lee et al., 2016) of

$$H_i^{E,\delta} : \theta_{i,O} - \theta_{i,R} = \delta$$

Which / how many CIs exclude 0 (after multiplicity adjustment)?

Selective z-test



Defining replicability

What do we estimate when we "estimate replicability?"

RP:P statistic 3: 83% of the effect size estimates declined from original to replication

Definition 3: Overall effect decline

- *What fraction of effect sizes declined by at least* 20%*?*
- Refers to *true* effect sizes, not point estimates

Answers question: "Do effects systematically attenuate in replications?"

Defining replicability

Definition 3: Overall effect decline

What fraction of effect sizes declined by at least 20%*?*

Did replication i show decline by at least $\rho \in [0, 1]$?

$$H_i^{D,\rho} : S_i \cdot \theta_{i,R} \geq S_i \cdot (1 - \rho)\theta_{i,O}$$

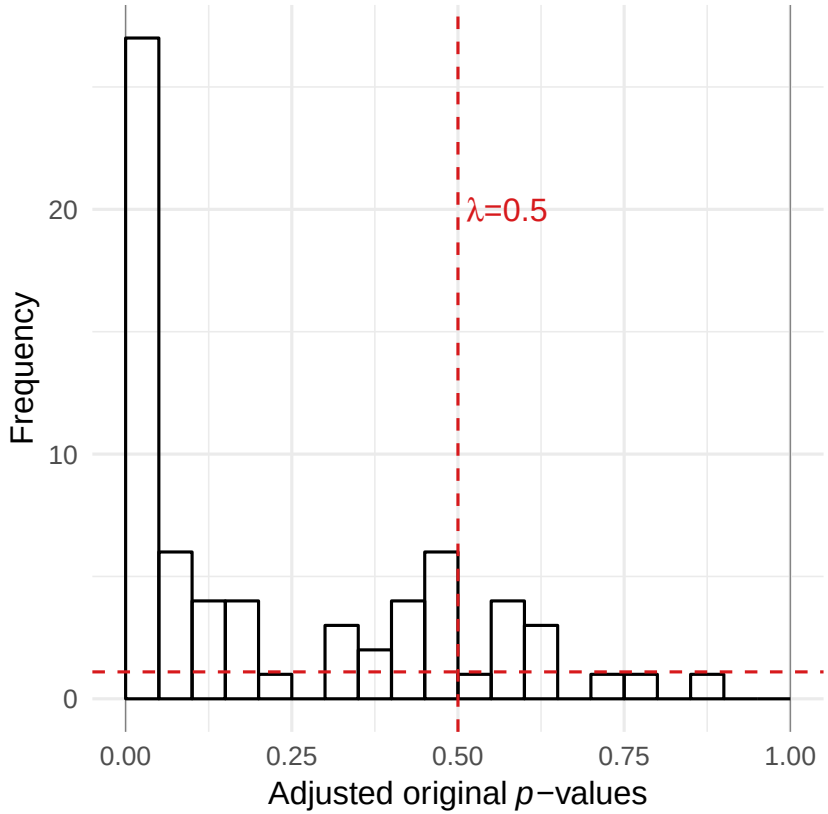
After conditioning on S_i, this is a linear hypothesis in $(\theta_{i,R}, \theta_{i,O})$

- Use Lee et al. (2016) test for individual $H_i^{D,\rho}$
- Aggregate to estimate / bound fraction that declined by ρ

Table of Contents

False directional claims

Work with adjusted p-values: conditionally, $p_{i,O}/0.05 \geq_{\text{st}} U[0,1]$, under $H_i^{S,O}$



NB: no spike near 1 ($p_{i,O} \approx 0.05$)

Inference on $\text{FDP}_{0.05}$

Adjusted p-value	$H_i^{S,O}$ true	$H_i^{S,O}$ false	Total
$20p_{i,O} \leq \lambda$	$*$	$*$	$*$
$20p_{i,O} > \lambda$	U	$*$	B
Total	$V_{0.05}$	$*$	$R_{0.05} = m$

Inferences based on:

$$B \geq U \geq_{\text{st}} \text{Binom}(V, 1-\lambda) \tag{2}$$

$\mathbb{E}B \geq (1-\lambda)V_{0.05}$ leads to estimator (Storey, 2002):

$$\widehat{\text{FDP}}_{0.05} = \frac{B}{R_{0.05}(1-\lambda)} = 2B/m \quad \text{if } \lambda = 1/2$$

(2) also gives UCB $V^*_{0.05}$, leads to UCB

$$\text{FDP}^*_{0.05} = V^*_{0.05}/R_{0.05}$$

Related questions

So far, asking about Type S error:

- Does S_i is correctly describe sign($\theta_{i,O}$)? (true orig. effect)

Two related questions:

- Does S_i correctly predict sign($\theta_{i,R}$)? (true repl. effect)
- Would Type S error be better if we'd used a different threshold, e.g. 0.005?

Requires slightly more subtle methods, similar in spirit

Results: False directional claims

Note:

- Estimate FDP $\approx 32\%$ for 0.05 threshold
- For $\alpha = 0.005$, estimate is 7%, UCB 18%
- But not clear that FDP for replications is improved
- Numbers overestimate Type S error ($\theta = 0.001$ similar to 0).

Estimates / CIs for false directional claims

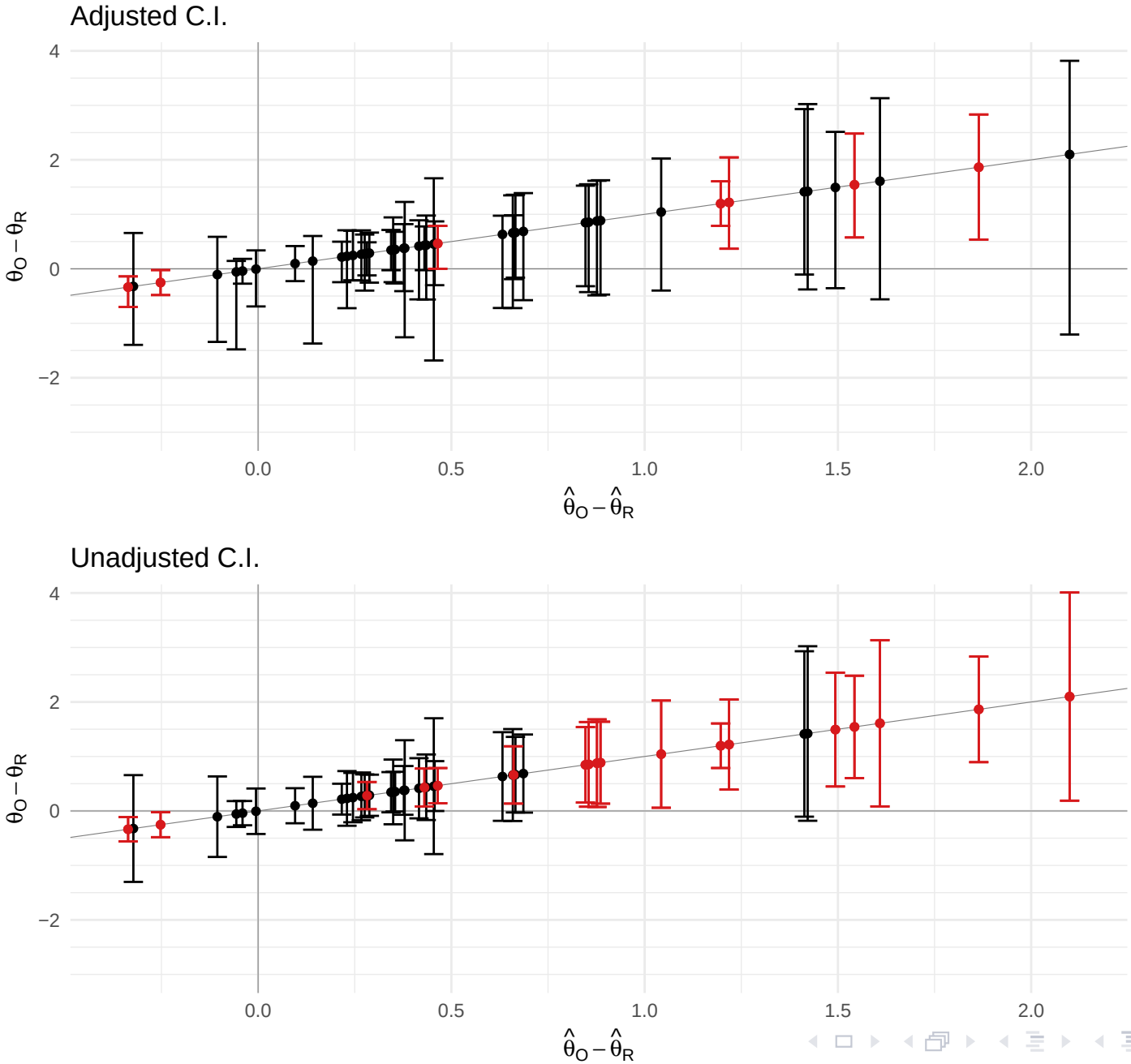
α	Orig. Est.	Orig. U.C.B.	Repl. Est.	Repl. U.C.B.
0.001	2%	9%	27%	55%
0.005	7%	18%	36%	61%
0.01	11%	22%	39%	61%
0.05	32%	47%	47%	63%

Results: Effect shift

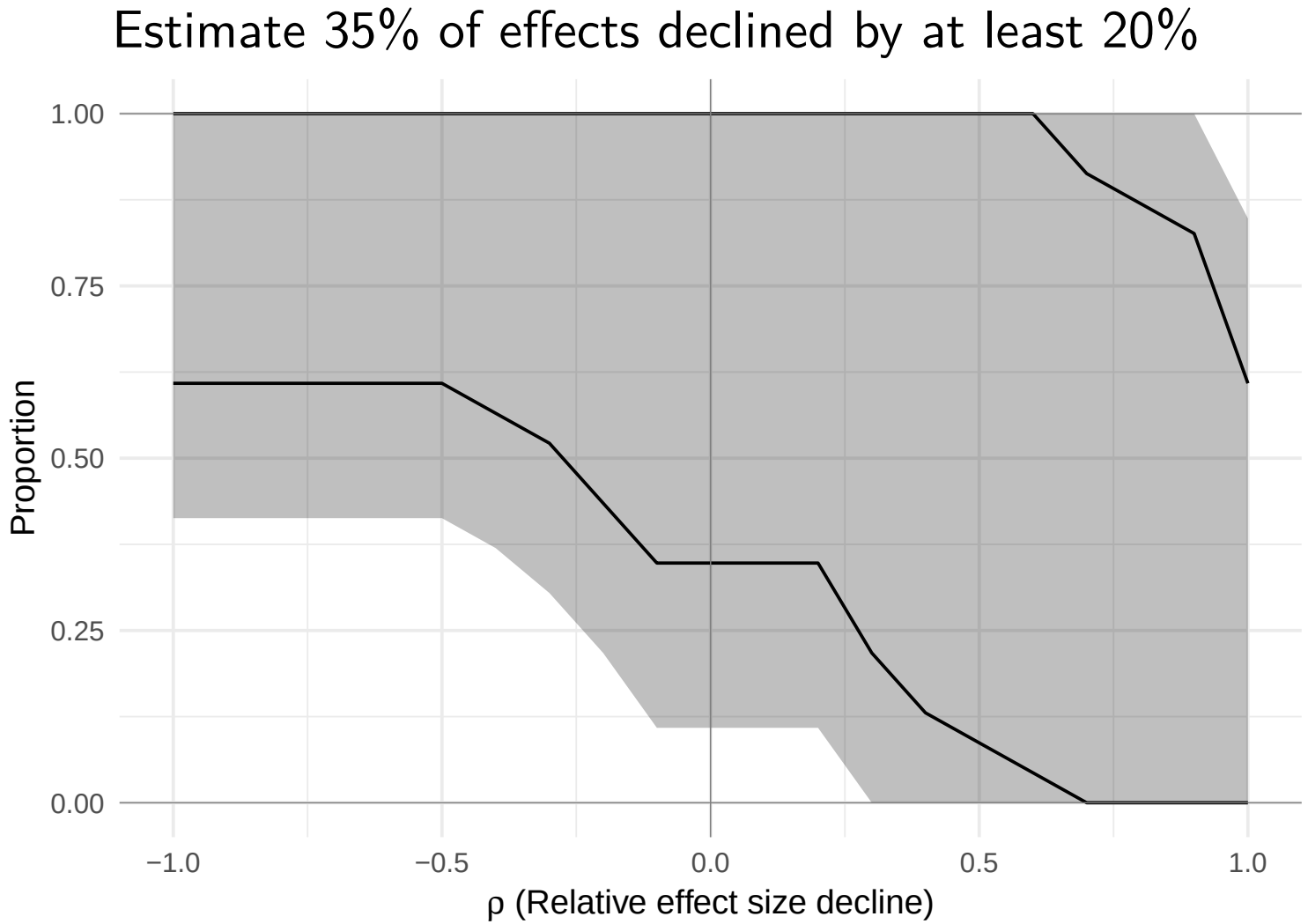
We construct 95% CIs for $\theta_{i,O} - \theta_{i,R}$:

- 15% of studies exclude 0 (exact replication), adjusting for selection
- 39% of studies rejected, otherwise
- 11% exclude 0 after $\text{BH}_{0.1}$ correction, 1 after Bonferroni

Results: Effect shift



Results: Effect decline



Takeaways

Selection bias is a powerful force

- Leads to many predictable pathologies
- Can't learn anything else without accounting for it
- Truncated Gaussian model opens many avenues for inference

Replicability has many possible meanings

- Precisely specifying estimand is essential for meaningful discussions

Rate of Type S errors in experimental psych is high: ($\approx 32\%$ of publ.?)

- Reducing threshold to 0.005 seems to improve FDP w.r.t. orig. effects
- Doesn't mean results will be replicable in new experiments

Evidence in a few studies that true effect sizes differ substantially

Some evidence of systematic effect decline (need more data)

Future work

- Preregistration
  - Better publication bias model
  - Less conservative estimate: information of nonsignificant studies are useful still!
- Higher powered design, e.g. Camerer et al. (2018); Klein et al. (2018)
- More formal criteria
- Clearer picture of the replicability crisis

Thanks!

References

The scientific method. *The Economist*, February 2016.

Joel Achenbach. Many scientific studies can't be replicated. That's a problem. *The Washington Post*, August 2015.

Christopher J Anderson, Štěpán Bahník, Michael Barnett-Cowan, Frank A Bosco, Jesse Chandler, Christopher R Chartier, Felix Cheung, Cody D Christopherson, Andreas Cordes, Edward J Cremata, Nicholas Della Penna, Vivien Estel, Anna Fedor, Stanka A Fitneva, Michael C Frank, James A Grange, Joshua K Hartshorne, Fred Hasselman, Felix Henninger, Marije van der Hulst, Kai J Jonas, Calvin K Lai, Carmel A Levitan, Jeremy K Miller, Katherine S Moore, Johannes M Meixner, Marcus R Munafò, Koen I Neijenhuijs, Gustav Nilsonne, Brian A Nosek, Franziska Plessow, Jason M Prenoveau, Ashley A Ricker, Kathleen Schmidt, Jeffrey R Spies, Stefan Stieger, Nina Strohminger, Gavin B Sullivan, Robbie C M van Aert, Marcel A L M van Assen, Wolf Vanpaemel, Michelangelo Vianello, Martin Voracek, and Kellylynn Zuni. Response to Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037c, March 2016.

References

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, Eric-Jan Wagenmakers, Richard A Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin F Camerer, David Cesarini, Christopher D Chambers, Merlise Clyde, Thomas D Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P Field, Malcolm Forster, Edward I George, Richard Gonzalez, Steven N Goodman, Edwin Green, Donald P Green, Anthony G Greenwald, Jarrod D Hadfield, Larry V Hedges, Leonhard Held, Teck-Hua Ho, Herbert Hoijtink, Daniel J Hruschka, Kosuke Imai, Guido W Imbens, John P A Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E Maxwell, Michael McCarthy, Don A Moore, Stephen L Morgan, Marcus R Munafò, Shinichi Nakagawa, Brendan Nyhan, Timothy H Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J Watts, Christopher Winship, Robert L Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2:6–10, January 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 57(1):289–300, 1995.

References

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 343:229–268, August 2018.

Benedict Carey. Many psychology findings not as strong as claimed, study says. *The New York Times*, page A1, August 2015.

William Fithian, Dennis L Sun, and Jonathan E Taylor. Optimal Inference After Model Selection. *arXiv*, October 2014.

Andrew Gelman and Francis Tuerlinckx. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037a, 2016a.

References

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. More on "Estimating the Reproducibility of Psychological Science", March 2016b. URL https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_post_publication_response.pdf.

Daniel T Gilbert, Gary King, Stephen Pettigrew, and Timothy D Wilson. A Response to the Reply to Our Technical Comment on "Estimating the Reproducibility of Psychological Science" , March 2016c. URL https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_response_to_osc_rebutal.pdf.

Ruth Heller, Yulia Golland, Rafael Malach, and Yoav Benjamini. Conjunction group analysis: an alternative to mixed/random effect analysis. *Neuroimage*, 37(4): 1178–1185, 2007.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams, Sinan Alper, Mark Aveyard, Jordan R Axt, Štěpán Bahník, Rishtee Batra, Mihály Berkics, Michael J Bernstein, Daniel Berry, Olga Bialobrzeska, Evans Binan, Konrad Bocian, Mark J Brandt, Robert Busching, Anna Cabak Rédei, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl L Carmichael, Francisco Ceric, David C Cicero, Jesse Chandler, Armand Chatard, Eva E Chen, Jen-Ho Chang, Winnee Cheong, Sharon Coen, Jennifer A Coleman, Brian Collisson, Morgan A Conway, Katherine S Corker, Paul G Curran, Fiery Cushman, Zubairu K Dagona, Ilker Dalgar,

References

Anna Dalla Rosa, William E David, Maaike de Brujin, Leander De Schutter, Thierry Devos, Canay Doğulu, Nerisa Dozo, Kristin Nicole Dukes, Yarrow Dunham, Kevin Durrheim, Charles R Ebersole, John E Edlund, Alexander Scott English, Anja Eller, Carolyn Finck, Natalia Frankowska, Miguel-Ángel Freyre, Mike Friedman, Elisa Maria Galliani, Joshua C Gandi, Tanuka Ghoshal, Steffen R Giessner, Tripat Gill, Timo Gnambs, Ángel Gómez, Roberto González, Jesse Graham, Jon E Grahe, Ivan Grahek, Eva G T Green, Kakul Hai, Matthew Haigh, Elizabeth L Haines, Michael P Hall, Marie E Heffernan, Joshua A Hicks, Petr Houdek, Jeffrey R Huntsinger, Ho Phi Huynh, Hans IJzerman, Yoel Inbar, Åse H Innes-Ker, William Jiménez-Leal, Melissa-Sue John, Jennifer A Joy-Gaba, Anna Kende, Roza G Kamiloğlu, Heather Barry Kappes, Serdar Karabati, Haruna Karick, Victor N Keller, Nicolas Kervyn, Goran Knežević, Carrie Kovacs, Lacy E Krueger, German Kurapov, Jamie Kurtz, Daniël Lakens, Ljiljana B Lazarević, Carmel A Levitan, Jr Neil A Lewis, Samuel Lins, Nikolette P Lipsey, Joy Losee, Esther Maassen, Angela T Maitner, Winfrida Malingumu, Robyn K Mallett, Saita A Marotta, Janko Međedović, Fernando Mena Pacheco, Taciano L Milfont, Wendy L Morris, Sean Murphy, Andriy Myachykov, Nick Neave, Koen Neijenhuijs, Anthony J Nelson, Félix Neto, Austin Lee Nichols, Aaron Ocampo, Susan L ODonnell, Elsie Ong, Malgorzata Osowiecka, Gábor Orosz, Grant Packard, Rolando Pérez-Sánchez, Boban Petrović, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Gabrielle Pogge, Monique M H Pollmann, Abraham M

References

Rutchick, Alexander Saeri, Patricio Saavedra, Erika Salomon, Kathleen Schmidt, Felix D Schönbrodt, Maciej B Sekerdej, David Sirlopú, Jeannie L M Skorinko, Michael A Smith, Vanessa Smith-Castro, Karin Smolders, Agata Sobkow, Walter Sowden, Manini Srivastava, Oskar K Sundfelt, Philipp Spachtholz, Troy G Steiner, Jeroen Stouten, Chris N H Street, Stephanie Szeto, Ewa Szumowska, Andrew Tang, Norbert Tanzer, Morgan Tear, Manuela Thomae, Jakub Traczyk, David Torres, Jordan Theriault, Joshua M Tybur, Adrienn Ujhelyi, Robbie C M van Aert, Marcel A L M van Assen, Paul A M van Lange, Marije van der Hulst, Anna Elisabeth van 't Veer, Alejandro Vásquez Echeverría, Leigh Ann Vaughn, Alexandra Vásquez, Luis Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid Voermans, Marek A Vranka, Marieke de Vries, Cheryl Welch, Aaron Wichman, Lisa A Williams, Michael Wood, Julie A Woodzicka, Marta K Wronska, Liane Young, John M Zelenski, Zhijia Zeng, and Brian A Nosek. Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. oct 2018. URL `https://psyarxiv.com/9654g/`.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

References

Brian A Nosek and Elizabeth Gilbert. Let's not mischaracterize replication studies: authors, March 2016. URL `https://retractionwatch.com/2016/03/07/lets-not-mischaracterize-replication-studies-authors/`.

Sanjay Srivastava. Evaluating a new critique of the reproducibility project, March 2016. URL `https://thehardestscience.com/2016/03/03/evaluating-a-new-critique-of-the-reproducibility-project/`.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 64(3):479–498, July 2002.