ExpEcon Methods:
Ethical & Unethical Research Practices

ECON 8877

P.J. Healy

First version thanks to Irfan Khan

Updated 2026-01-28

Replication Crises & Data Colada

- Replication crisis in psychology & social science: mid-2010s
  - Concerns had been floating around since the 1960s...
  - Social Psych hit especially hard
- Replication projects re-running existing experiments
  - Nosek et al. (2015): Only 36% of results replicated!!
    - Social psych: 25%
    - Cognitive psych: 50%
  - Camerer et al. (2016): Experimental economics papers
    - 11 of 18 (61%) replicated
- Data Colada blog identified systemic problems
  - Co-authored by data sleuths, notably Uri Simonsohn
  - Identified outright fraud by several famous economists

Dishonesty in Research

There are two widely recognized types of research-driven publication biases

- Selection Problems: The “file drawer effect”
  - Studies with nonsignificant effects have lower publication rates
  - Version of this for ExpEcon: “g-hacking” (game hacking)
- Inflation Bias: “p-hacking” or “selective reporting”
  - Data analysis practices that lead to false positives
  - Strategic reporting of favorable specifications/results

Do these only come from maliciously fraudulent researchers? NO!

File Drawer Bias

- Assuming the Null is true, if 100 studies are performed, 5 of them should yield statistically significant results
- If only these 5 are sent in for publication, then the community may believe that these are indicative of the true effect, while in fact they are not
- Many researchers have huge budgets, and can carry out many studies, and put the ones that do not produce significant results in the file drawer
- How to correct?
  1. Replication by self or others.
     - Currently: not very lucrative
     - Journals should publish null results & replications (JESA)
  2. Requiring robustness checks (but that's unfair)
- Discuss: does pre-registration/pre-analysis plan fix this?

P-Hacking

"P-Hacking": unethical techniques to try to get a significant result

"False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" by Simmons, Nelson, and Simonsohn

- Researchers have a lot of flexibility in their analyses:
  A. Choosing the best dependent variables/outcome measures
  B. Adding to the sample size if *p*-value is "close"
  C. Adding/removing covariates (gender, IQ, etc.)
  D. Discarding "outliers" or even treatments ex-post

They simulate some of these "tricks" for a hypothetical study:

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Illustration: Combining Pilots With Data

A simulation:

1. Run a pilot with n_p subjects
   - Generate n_p observations of X_i^p and Y_i^p from $N(0, 1)$
   - Run a t-test on X^p vs Y^p
   - $\approx 5\%$ will (wrongly) reject H_0

2a. If pilot fails to reject, stop the project! It's a dud

2b. If pilot rejects, run full sample with n_s subjects
   - Generate n_s observations of X_i^s and Y_i^s from $N(0, 1)$
   - Two options:

Ethical: Analyze new samples only: X^s vs. Y^s. Throw away the pilot.

Unethical: Analyze combined samples: (X^p, X^s) vs. (Y^p, Y^s)

Simulation: Repeat this 10,000 times, look at rejection rates

How bad will it be?

Simulation Results

Simulation #:	1	2	3
Pilot n_p:	100	100	500
Sample n_s:	100	500	100
# Simulations:	10,000	10,000	10,000
% where Pilot rejects:	0.0483	0.0511	0.0463
# Continued Studies:	483	511	463
% Reject (New Data Only):	0.056	0.053	0.048
% Reject (Combined Data):	0.354	0.160	0.631

You're selectively picking only pilots with false positives!

What about checking your data?

Entirely hypothetical question:

- Suppose you're a nervous young researcher
- Maybe your experiment software has a bug!!
- So, you run 50 subjects on Prolific to make sure it works
- If it looks okay, you run 300 more. If not, stop and fix.

Is this a problem? (discuss)

What about checking your data?

Entirely hypothetical question:

- Suppose you're a nervous young researcher
- Maybe your experiment software has a bug!!
- So, you run 50 subjects on Prolific to make sure it works
- If it looks okay, you run 300 more. If not, stop and fix.

Is this a problem? (discuss)

No, as long as either

1. you throw away the first 50 subjects, or
2. your stop/go decision doesn't depend on the statistical test result (just on "data quality")

But wait: is your 50-person pilot really all that well-powered??
Also see discussion of "design hacking" later...

Unethical Sequential Sampling

More generally, consider this (unethical) sampling algorithm:
Parameters:

- Initial sample size: n. Max you can afford: $\bar{n} > n$.
- Keep adding subjects as long as p-value is in $[0.05, \bar{p}]$
- Subjects added at each step: n_a

Algorithm:

1. Collect n initial observations each of X and Y
    - Suppose the null is true. e.g. $X, Y \sim N(0, 1)$
2. Run test.
    - If $p < 0.05$, stop. You win! H_0 is rejected! Publish!
    - If $p > \bar{p}$, stop. It's hopeless. You lose. File drawer.
    - Otherwise, continue to next step:
3. Add another n_a observations each of X and Y
    - If $n + n_a > \bar{n}$, stop. You ran out of money. File drawer.
    - Otherwise, run test (return to step 2),

How bad can it be?

Unethical Sequential Sampling

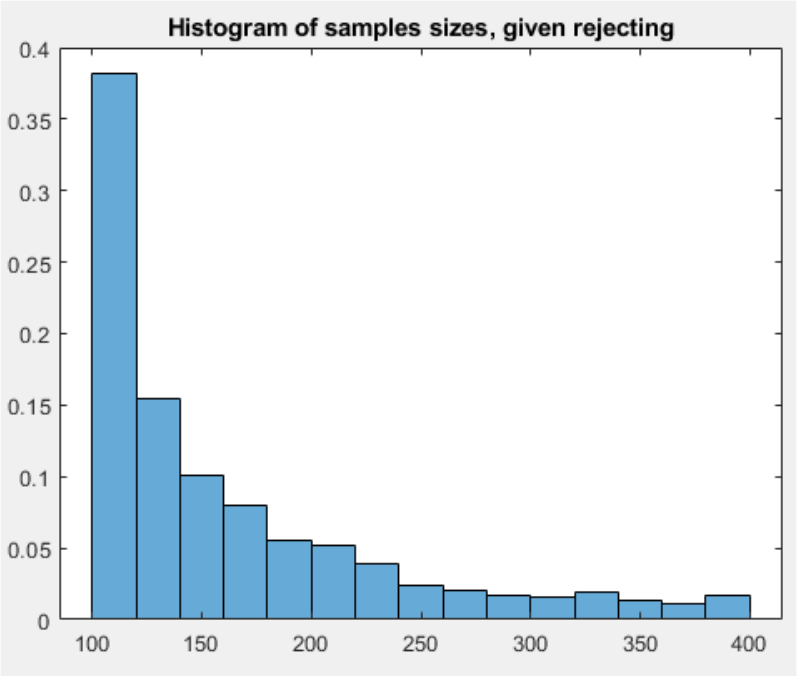
Rejection frequencies, varying give-up $\bar{p}$

Simulation #:	1	2	3
Initial n:	100	100	100
Added n_a:	10	20	20
Max $\bar{n}$:	200	200	400
$\bar{p} = 0.10$	0.0665	0.0666	0.0649
$\bar{p} = 0.15$	0.0785	0.0779	0.0744
$\bar{p} = 0.20$	0.0914	0.0843	0.0907
$\bar{p} = 0.30$	0.1011	0.0991	0.1023
$\bar{p} = 0.50$	0.1172	0.1117	0.1378

1. Increasing $\bar{p}$ (give-up threshold) $\Rightarrow$ more false positives
2. Increasing $n_a \Rightarrow$ fewer tries $\Rightarrow$ slightly fewer false positives
3. Increasing $\bar{n}$ (budget) $\Rightarrow$ depends on $\bar{p}$

Unethical Sequential Sampling

How much do you spend? ($n = 100$, $n_a = 20$, $\bar{n} = 400$, $\bar{p} = 0.50$)

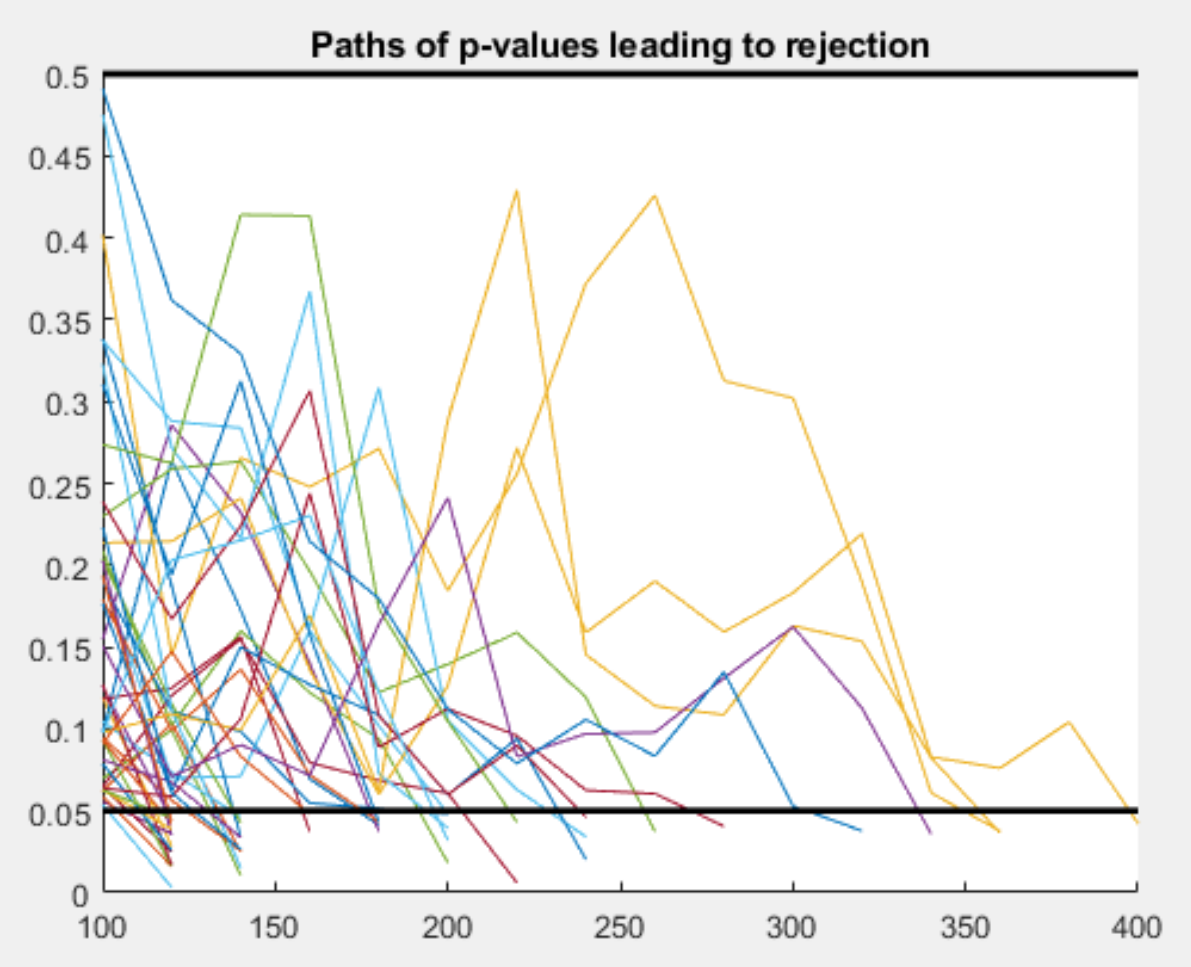


Avg: 154 subjects. Median: 120 subjects

Quit because $p > 0.50$: 83%
Quit because $n > 400$: 3.6% $\leftarrow$ budget not binding

Unethical Sequential Sampling

Paths of p-values that led to rejection:
($n = 100$, $n_a = 20$, $\bar{n} = 400$, $\bar{p} = 0.50$)



Ethical Sequential Sampling

- There are ethical sequential sampling procedures...
- Wald's Sequential Probability Ratio Test
  - Requires 2 specific, parameterized hypotheses
  - Ex: H_0: $N(0, 1)$ vs H_1: $N(1, 1)$
  - Let $p(x_i|0)$ and $p(x_i|1)$ be likelihoods of x_i under each
  - Likelihood ratio of H_1 for data vector $x = (x_1, \ldots, x_n)$:

$$\frac{p(x_1|1)\ p(x_2|1)\ \cdots\ p(x_n|1)}{p(x_1|0)\ p(x_2|0)\ \cdots\ p(x_n|0)} \rightarrow \sum_i log\left(\frac{p(x_i|1)}{p(x_i|0)}\right)$$

  - Under H_0, compare to test error ratio:

$$\frac{p(x_1|1)\ p(x_2|1)\ \cdots\ p(x_n|1)}{p(x_1|0)\ p(x_2|0)\ \cdots\ p(x_n|0)} = \frac{\beta}{1-\alpha} \rightarrow \log\left(\frac{\beta}{1-\alpha}\right)$$

  - Collect data *sequentially*, monitoring the total log-likelihood ratio
  - If it falls below $a = \log(\beta/(1-\alpha))$, accept H_0
  - If it rises above $b = \log((1-\beta)/\alpha)$, accept H_1
- $\exists$ a sequential test for a single hypothesis?

Table 2. Simple Solution to the Problem of False-Positive Publications

Requirements for authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Guidelines for reviewers

1. Reviewers should ensure that authors follow the requirements.
2. Reviewers should be more tolerant of imperfections in results.
3. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
4. If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.

Solutions

1. Pre-Registration (AEA RCT Registry, e.g.)
    - Basic plan of your research
    - Design, sample size, clusters, dates
    - Problem: can be vague, so not very constraining
2. Pre-Analysis Plans (PAPs)
    - Much more detailed
    - Instructions, regressions, hyp. tests, etc.
    - You can deviate, but would need to document it
    - Problem: referees might punish deviations
    - Problem: might kill scientific discovery
3. Registered Reports
    - Journal decides *before* you collect data
    - Problem: Very few journals do this (JPE:Micro, JESA)
    - Problem: If results are great you'll want to renege!
    - Problem: Editors become advisors

My Recommendation

My (current) recommendation:

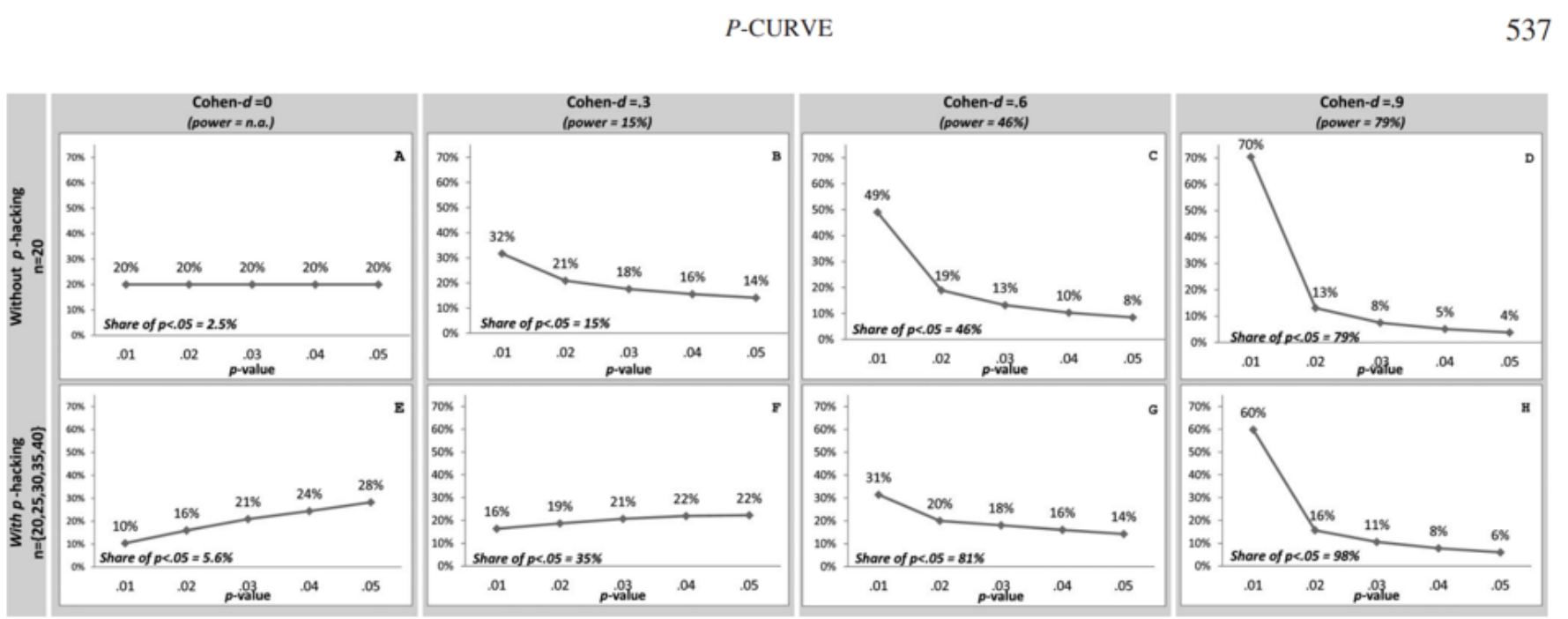
1. Might as well create a pre-analysis plan.
2. One step further: generate simulated data *before* running
    2.1 Figure out your power
    2.2 Nail down the tests you want
    2.3 Make sure your tests actually answer the right question!!!
    2.4 Now your code is ready! Gather the data and hit “play”
3. In the paper, be honest about exploratory results
    3.1 Cross your fingers that referees are willing to listen

Evidence for P-hacking

How to identify P-hacking?

- “P-Curve: A Key to the File-Drawer” by Simonsohn, Nelson, and Simmons
- Look at the distribution of p-values in a literature
- What should the distribution look like below 0.05??
  - Red flag: lots of values just below 0.50
    - That shouldn’t happen naturally!

P-Curve under certain distributions



Look for “uphill” or “flattened” curves

A demonstration

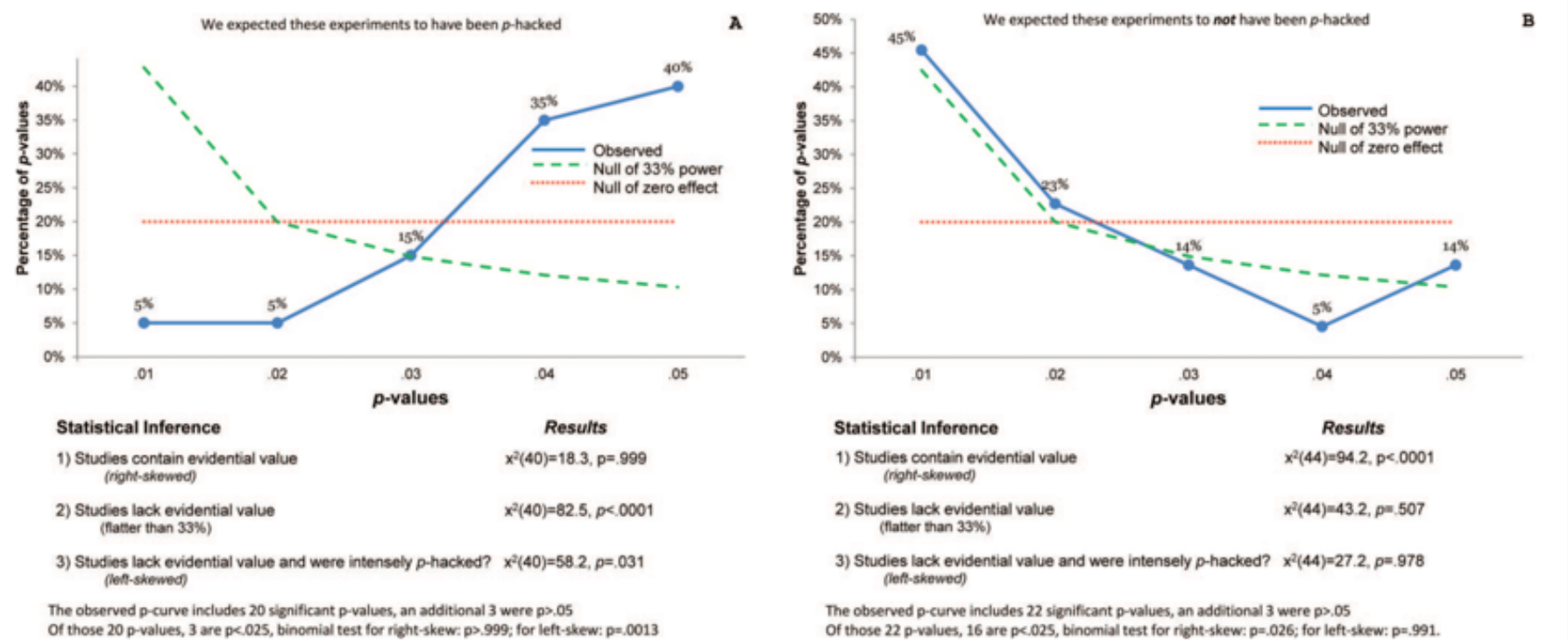


Figure 3. *P*-curves for *Journal of Personality and Social Psychology (JPSP)* studies suspected to have been *p*-hacked (A) and not *p*-hacked (B). Graphs depict *p*-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 *JPSP* studies that only report statistical results from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 *JPSP* studies reported in articles whose full text does not include keywords that we suspected could indicate *p*-hacking (e.g., *exclude*, *covariate*).

Red flag: papers that add controls when treatment was random

Specification Curve Analysis

Simonsohn, Simmons, Lennon (2020)

- Report all results of all sensible specifications. Meaning:
  1. a sensible test of the research question,
  2. expected to be statistically valid, and
  3. not redundant with the other tests reported.
- Similar to applied micro's table of regressions

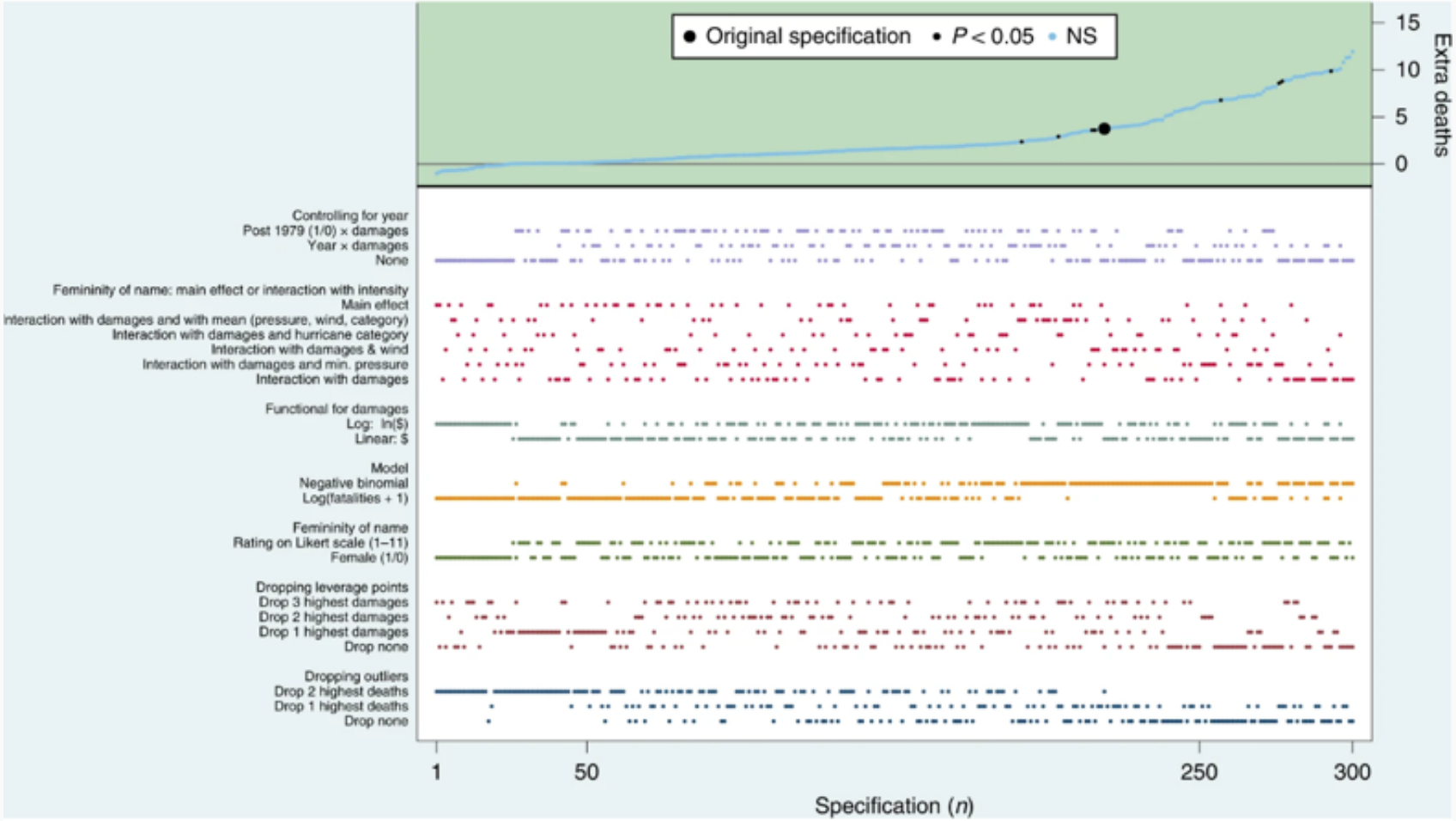
Specification Curve

- Your regression specification:

$$y = F(x; Z) + \epsilon$$

- Lots of degrees of freedom!
  - Different y (wealth, education...)
  - Different F (linear, polynomial...)
  - Different x (treatments, covariates...)
  - Different Z (gender, race, education...)
  - You can easily generate 100+ specifications

Example: “Hurricanes with female names cause more damage”



Each dot in the top panel (green area) depicts the marginal effect, estimated at sample means, of a hurricane having a female rather than male name; the dots vertically aligned below (white area) indicate the analytical decisions behind those estimates. A total of 1,728 specifications were estimated; to facilitate visual inspection, the figure depicts the 50 highest and lowest point estimates and a random subset of 200 additional ones, but the inferential statistics for specification curve analysis include all 1,728 specifications. NS, not significant.

Top: Marginal effects of female name on extra deaths.

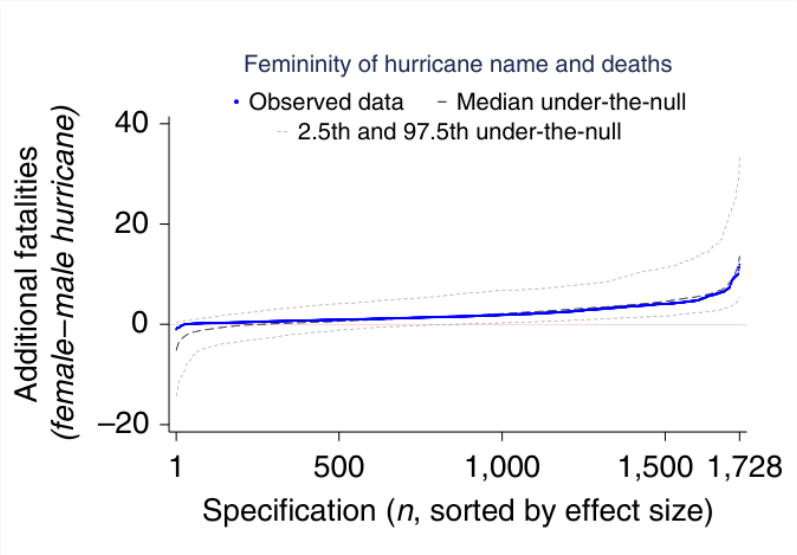
Height: estimated effect size. Black dot: $p < 0.05$

Bottom: dots show specification choices for points on the line above

How To Analyze This?

Bootstrapping!

1. Reshuffle the hurricane names, but nothing else (null is true)
2. Run the specification curve on the bootstrapped sample
3. Repeat many times, plot each curve



Median effect size (across *all* specifications) using *true* names: 1.56.
% Bootstrapped medians $>$ 1.56? 0.536 of them $\leftarrow$ p-value
(Can use "% significant specifications" instead of median effect size)

“Design Hacking”

“Design Hacking” (my term) or “Game Hacking” (Muriel’s term)
File-drawer bias can happen “within” a project as well:

- Try one design, throw it on Prolific, get a null result
- Tweak your design, keep trying, until finally you reject the null
- Collect a fresh new full sample using the design that worked

Is it problematic?? Discuss.

“Design Hacking”

“Design Hacking” (my term) or “Game Hacking” (Muriel’s term)
File-drawer bias can happen “within” a project as well:

- Try one design, throw it on Prolific, get a null result
- Tweak your design, keep trying, until finally you reject the null
- Collect a fresh new full sample using the design that worked

Is it problematic?? Discuss.

1. Fresh new full sample $\Rightarrow$ not p-hacking
2. But you now **know** that your design isn’t robust!!
    - Are you comfortable if someone tries to replicate your result?
    - Doing this often *should* be a bad career choice...
    - Sadly: literature gets trapped in one specific design
      See: Ellsberg and Allais paradoxes

Solution (again): replications and robustness checks!!

Coffman & Dreber's Chapter

"Running Replicable Experiments" for new Handbook (Yariv & Snowberg)

- Idea: suppose your paper will be replicated. Great!
- But wait... are you scared now???
- This chapter: how to make sure your results replicate

Coffman & Dreber

They summarize it with two main guidelines:

1. Be very clear and comprehensive in describing everything you did.
2. Be rigorous and self-critical in your own work
    - If someone replicates your work, you can be confident!

Coffman & Dreber

Useful terms:

Reproducibility: Can others reproduce what you did?

- **Computational reproducibility** Can others get the same results using your data & code?
- **Procedural reproducibility:** From your write-up and materials, can someone rerun the same experiment without your help?

Replicability: Will others get the same results as you?

- **Direct replication:** Same instructions, type of subjects, etc.
- **Conceptual/robust replication:** Test of the same hypotheses in a slightly different setting

Coffman & Dreber

Should you pilot? Might that improve replicability?

Three types of pilot studies:

1. Testing software, subject confusion, etc.
2. Measuring means & variances for power calculations
3. Seeing if the hypothesis is right

Coffman & Dreber

Should you pilot? Might that improve replicability?

Three types of pilot studies:

1. Testing software, subject confusion, etc.
2. Measuring means & variances for power calculations
3. Seeing if the hypothesis is right

(1) is good, (2) is iffy (probably wasteful), (3) is bad

Coffman & Dreber

Should the profession require pre-analysis/pre-registration?

- **Single-paper view**: conclusive evidence re: a hypothesis can/should happen within a single paper
  - That one paper better be extremely solid! Require PAP!
  - But, are we punishing exploration?
  - And incentivizing design-hacking?
- **Literature-level view**: a body of literature collectively tests a hypothesis. Process of "explore and replicate."
  - Much more room for exploration. Replication replaces PAP
  - But then, can we trust any given paper?
  - And are we really replicating enough??

Two more thoughts:

1. Often, my control was your treatment. Free replication!
2. Theory testing: are there really that many dof?

Coffman & Dreber

Should **you** choose to do a Pre-analysis plan??
Pros:

1. Eliminates your own degrees of freedom. Ties your hands.
2. Referees & editors are looking for it more and more
3. Helpful to be more thoughtful about your research
4. Lets you do things like exclude outliers or confused subjects in a scientifically-credible way
5. Forces coauthors to agree *ex-ante* on the right analyses

Cons:

1. More likely to get a null result! (Which is honest...)
2. If you have to deviate (and you will), those results can be heavily discounted by referees as “exploratory”
3. Might make you less adventurous as a researcher

Coffman & Dreber

Power calculations

- Not just for the reader... they boost your own confidence!
- Extra important if a null result would be interesting

Inputs:

1. Pilots? Probably a waste
2. Your budget, of course
3. Literature! Existing effect sizes (incl. SE's)
4. Ask other experts?

Coffman & Dreber

Replication Packages

- Zip file of all original data and code (with readme) to guarantee computational reproducibility
  - Some journals actually have a data editor that reproduces your results before you can publish!
- Experiments: include instructions and screenshots.
- Comment code well *as you write it*
  - Imagine someone is reading your code as you type it

A Quick Discussion on Deception

Deception

Another ethical issue: Deception

- Estimates: ∼50% of papers in social psych
- It's (informally) banned in econ
  - Subject trust is a public good across experiments
  - We need them to believe our instructions!
- What counts?
  - Lying to subjects
  - Surprise treatments/questions?
  - Hiding information from subjects???
- Blatant deception unlikely to publish in Econ
  - Vernon credits Sidney Siegel for this norm
  - Really implemented by Plott and Smith, others
- There are gray areas... see below

Deception: Some Evidence

Jamison Karlan & Schechter (2008)

- Control: play trust games against human opponents
- Treatment: same, but opponents are actually computers
  - Programmed to play the same as the humans
  - Deception *was* revealed at the end
- All subjects recruited for a 2nd experiment 2–3 weeks later
  - Variety of different games
- Results:
  1. 2nd-experiment participation marginally lower among the deceived
     1.1 Effect was significant & large for women
  2. 2nd-experiment participation not correlated with behavior
  3. Deceived: more likely to be multiple-switchers in MPL
  4. Deceived: more variance in responses

Lesson: deception scares people away (differentially!) and maybe causes them to take future experiments “less seriously”

Deception

Charness, Samek, and van Den Van (2022)

- Survey of experimental econ researchers
- What counts as deception?
- 788 of 1554 responded
- Also surveyed experiment participants

Scenario	Researcher text
S1: Subgroup re-match	In a multi-period experiment, the experimenter tells the participants that they will be randomly matched every period, but in fact the participants are only re-matched (for statistical purposes) within a subgroup of the participants
S2: Surprise re-start	Participants in an experiment are told that there will be 10 periods in the session, but are then told that there will be another 10 periods (a surprise re-start)
S3: Non-representative sample	The experimenter tells the participants the average value of the choices or beliefs of "a sample of the other participants", but doesn't mention that this is not a representative sample (and states other averages to other participants)
S4: Unexpected data use	The experimenter uses participant responses in a way that is not revealed to the participant: for example, (1) participants are incentivized to predict behavior of other people, but are not told that these predictions will be shown to others, or (2) participant data from one part of the experiment is used to sort participants into groups in another part of the experiment
S5: Confederates	The experimenter uses either confederates or computers that do not operate of their own volition, but instead behave as scripted by the experimenter. The experimenter does not tell subjects that confederates or computers are involved in the experiment
S6: Unknown/unpaid participation	The experimenter conducts a field experiment that encourages people to put forth (unpaid) effort or take action, but does not inform the participants that they are in an experiment
S7: Misinterpretation	The experimenter relies upon the assumption that participants will misinterpret the instructions [e.g., using the term "random" when the probabilities are actually 75% and 25% and when it is essential that they believe that this was truly random (i.e., 50%)]

Scenario	Deceptive (1–7)	Negative (1–7)	Appropriate (1–7)	Useful (1–7)
Unexpected data use	3.18	2.94	5.19	4.96
	(0.07)	(0.07)	(0.06)	(0.06)
Subgroup re-match	3.20	3.01	5.00	4.64
	(0.07)	(0.08)	(0.07)	(0.07)
Unknown/unpaid participation	3.23	2.85	5.25	
	(0.08)	(0.07)	(0.07)	
Non-representative sample	3.76	3.42	4.76	4.40
	(0.07)	(0.07)	(0.07)	(0.07)
Surprise re-start	3.88	3.45	4.75	4.41
	(0.07)	(0.07)	(0.07)	(0.07)
Misinterpretation	4.78	4.58	3.70	
	(0.07)	(0.07)	(0.07)	
Confederates	5.33	4.79	3.88	4.07
	(0.07)	(0.07)	(0.07)	(0.07)
Total	3.91	3.58	4.65	4.50
	(0.03)	(0.03)	(0.03)	(0.03)

Mean ratings. Items are rated on a 7-point scale, ranging from 1 (“not at all”) to 7 (“extremely”).

Questions asked:

1. How **deceptive** is it? 2. Would you feel **negative** as a referee?
3. How **appropriate** is it if ∄ alternative? 4. How **useful** is it?

My View

What do you think?

My View

My view:

- All that matters is whether subjects will believe the instructions next time they come to an experiment
  - This is a public good!
  - Ethical issues matter, but this conservative approach covers them
- My assumption: Likelihood that they care/notice is driven by likelihood that they regret their former actions
  - Example: Testing Gang-of-Four with a surprise restart
  - “Regret-inducing surprise”
  - “Regret-free” deception *might* be okay, but still risky!
- Isn’t it okay if they don’t find out?
  - How sure are you? What if they talk?
  - What if they read our papers?
- I think it’s rare that you *must* use deception

Experimenter Demand Effects

Smaller but pervasive issue: Experimenter Demand Effects

- Altering choices through framing/display
  - Example: Preference for Randomization
- Or, making it obvious what's the research question
  - Ex: Gender study, only ask about gender
- Directional effect may be unclear!
- Raises deeper questions about:
  1. What is a preference? Depends on framing?
  2. What does it mean to have "external validity"?
-

Can We Reduce Experimenter Demand Effects?

- Incentives: Vernon's "Dominance"
  - Camerer: larger stakes reduce noise
- Neutral framing/instructions
  - But isn't "neutral" just another frame??
- Reducing interaction with the experimenter
  - Read-alone instructions? Video?
- My view: every frame alters preferences.
  - There is no "neutral frame" or "true preference"
  - So just document the framing you used
  - Future researchers can test robustness

The de Quidt et al (2017) Method

de Quidt et al. (2017)
Example: effect of incentives on effort

1. Run original design as planned.
    - Control (0): no pay
    - Treatment (1): piece rate pay
    - Let the mean actions be $a^o(0)$ and $a^o(1)$
2. Run a new copy, but with a "strongly positive" demand
    - "You would be doing us a favor if you work hard"
    - Let mean actions be $a^+(0)$ and $a^+(1)$
3. Run a "strongly negative" demand experiment
    - "You would be doing us a favor if you are lazy"
    - Let mean actions be $a^-(0)$ and $a^-(1)$
4. Compare treatment effects
    - Original treatment effect: $a^o(1) - a^o(0)$
    - Lower bound on treatment effect: $a^-(1) - a^+(0)$

The de Quidt et al (2017) Method

Another usage:

- If $a^+ \approx a^o$ or $a^- \approx a^o$ then no big deal!
- Usually prior expectation of direction ($+$ or $-$)