



BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Implementation
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

Registration and Pre-Analysis Plans

Making research more transparent and reproducible

Garret Christensen¹

¹UC Berkeley:

Berkeley Initiative for Transparency in the Social Sciences
Berkeley Institute for Data Science

MCA Zambia, March 2016

Slides available online at <https://osf.io/m8ey6/>

Overview

- Publication bias is a problem.
- Registration can help with publication bias.
- P-hacking/specification searching is a problem.
- Pre-analysis plans can help with specification searching.
- What should we include in our PAP, and where should we post it?

Existence of the problem:

- Effect sizes diminish with sample size (Gerber, Green, Nickerson 2001)
- There is a higher fraction of rejected hypothesis tests in social compared to hard sciences (Fanelli 2010).
- Published null results are disappearing over time, in all disciplines (Fanelli 2011).
- Data on the complete set of experiments run shows strong results are 40pp more likely to be published, and 60pp more likely to be written up. The file drawer problem is large. (Franco, Malhotra, Simonovits 2014)

JELLY BEANS
CAUSE ACNE!

SCIENTISTS!
INVESTIGATE!

BUT WE'RE
PLAYING
MINECRAFT!
... FINE.



WE FOUND NO
LINK BETWEEN
JELLY BEANS AND
ACNE ($p > 0.05$).



THAT SETTLES THAT.

I HEAR IT'S ONLY
A CERTAIN COLOR
THAT CAUSES IT.

SCIENTISTS!

BUT
MINECRAFT!



WE FOUND NO
LINK BETWEEN
PURPLE JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
BROWN JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
PINK JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
BLUE JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
TEAL JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
SALMON JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
RED JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
TURQUOISE JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
MAGENTA JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
YELLOW JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
GREY JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
TAN JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
CYAN JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND A
LINK BETWEEN
GREEN JELLY
BEANS AND ACNE
($p < 0.05$).



WE FOUND NO
LINK BETWEEN
MAGENTA JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
BESIDE JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
LAWA JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
BLACK JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
PINK JELLY
BEANS AND ACNE
($p > 0.05$).



WE FOUND NO
LINK BETWEEN
ORANGE JELLY
BEANS AND ACNE
($p > 0.05$).




News

**GREEN JELLY
BEANS LINKED
TO ACNE!**

95% CONFIDENCE

**ONLY 5% CHANCE
OF COINCIDENCE!**

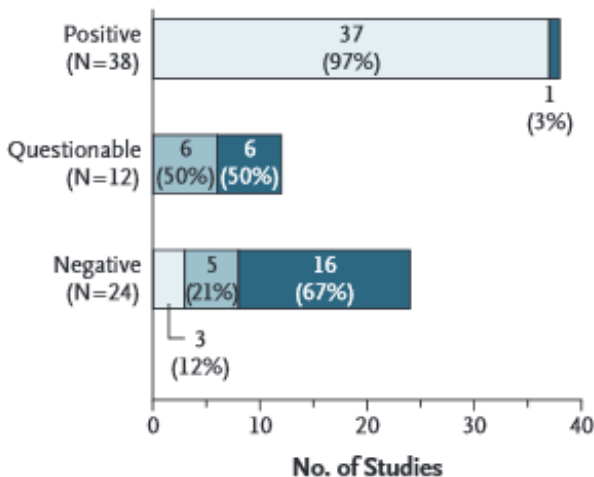
SCIENTISTS...



- Published, agrees with FDA decision
- Published, conflicts with FDA decision
- Not published

A Studies (N=74)

FDA Decision





Publication Bias

Implementation
Registration
and
Pre-Analysis
Plans
Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

If we only write up/publish significant results, and we have no record of all the insignificant results, we have no way to tell if our 'significant' results are real, or if they're the 5% we should expect due to noise.



BITSS

BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Publication Bias—Statistical Correction

Implementation:
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

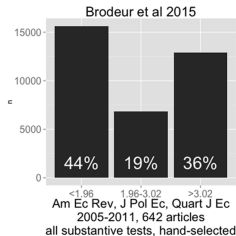
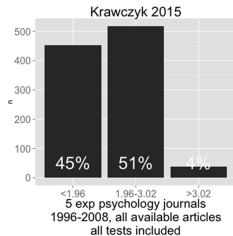
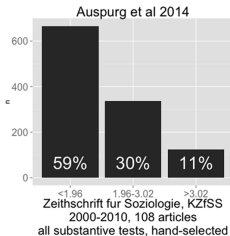
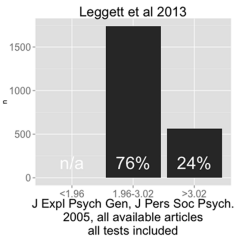
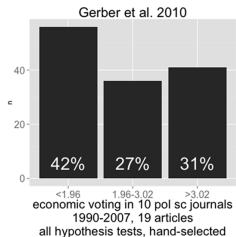
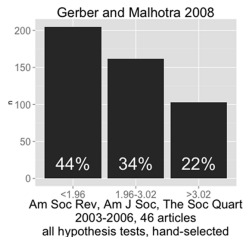
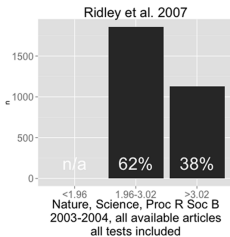
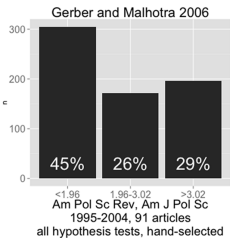
P-Hacking

Pre-Analysis
Plan

Conclusion

Under simple models of publication bias, you can restore the original type I error rate by adjusting the test statistic cutoff. About 30% of published tests fall between the unadjusted cutoff (1.96) and the adjusted cutoff (3.02).

However, this method breaks down if it were to be widely applied. (McCrary, Christensen, Fanelli 2016)



- Publicly stating all research you will do, and what hypotheses you will test, prospectively.
- Store this statement in a public registry.
- Near universal adoption in medical RCTs. Top journals (ICMJE) won't publish if it's not registered.
- Largest: <http://clinicaltrials.gov>
- Even better if registry requires registering outcomes after study. Currently limited, and poor compliance (Anderson et al. 2015) but NIH is moving on this. [Link](#)

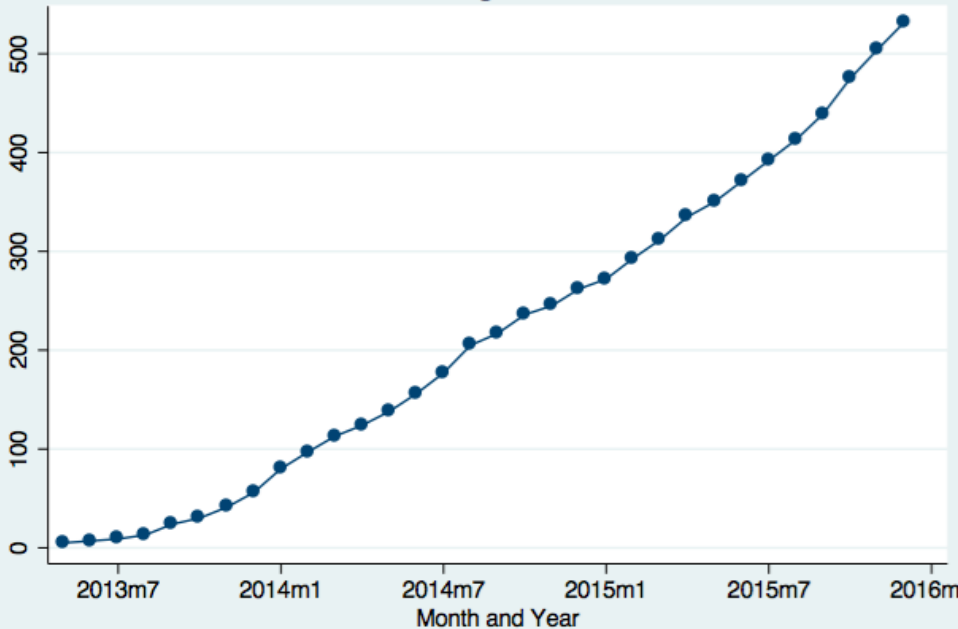
- Newer to social sciences, but good locations for several fields.

- AEA registry, currently only for RCTs.

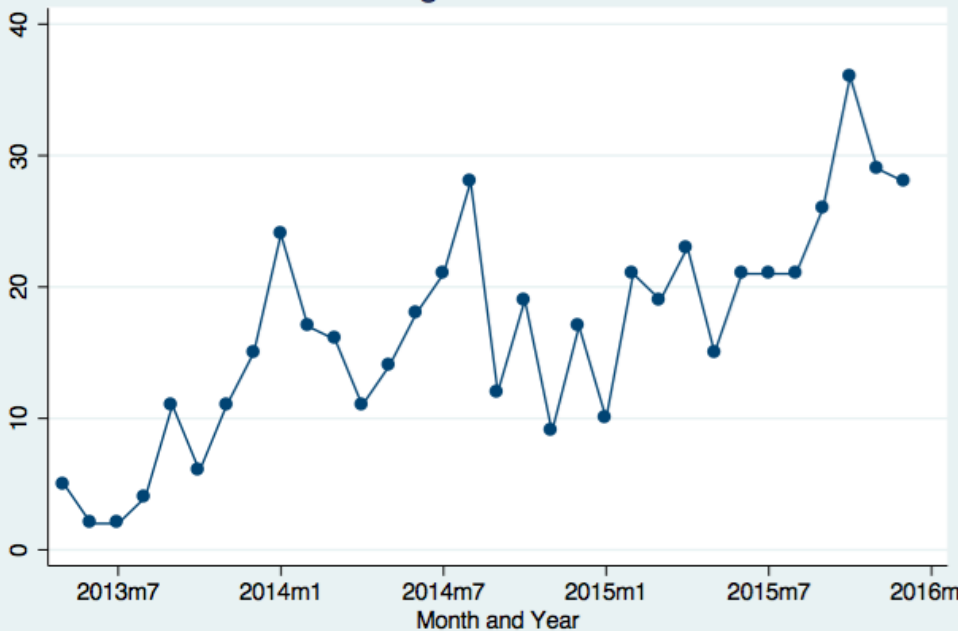
<http://socialscienceregistry.org>

- “J-PAL supports the American Economic Association’s (AEA) registry for randomized controlled trials in economics (<http://socialscienceregistry.org>). It is a free, easy-to-use database that makes access to trial results more transparent, aims to address the growing number of requests for registration by funders and referees, and helps solve the problem of publication bias by providing a single place where all trials are registered in advance of their start. As of May 31, 2015, the AEA Registry had a total of 379 registered controlled trials in 70 different countries. See how the registry continues to grow below.”

Total AEA Trial Registrations over Time



New AEA Registrations Each Month





Registrations-Social Sciences

BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Implementation
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

- EGAP registry

<http://egap.org/design-registration>

- 3ie registry, for developing country evaluations.

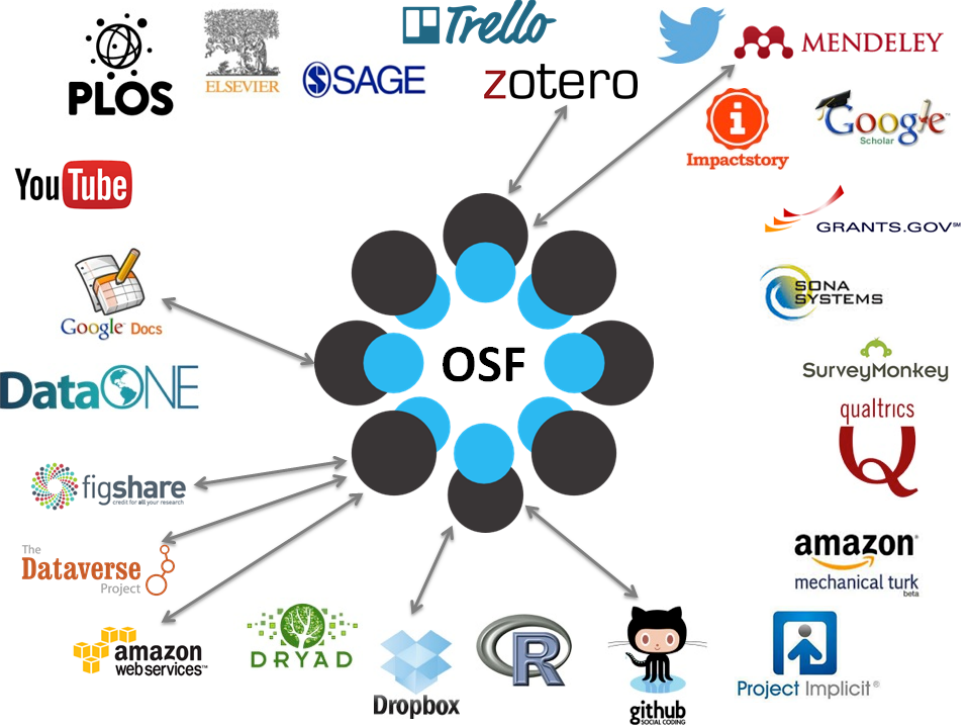
<http://ridie.3ieimpact.org>

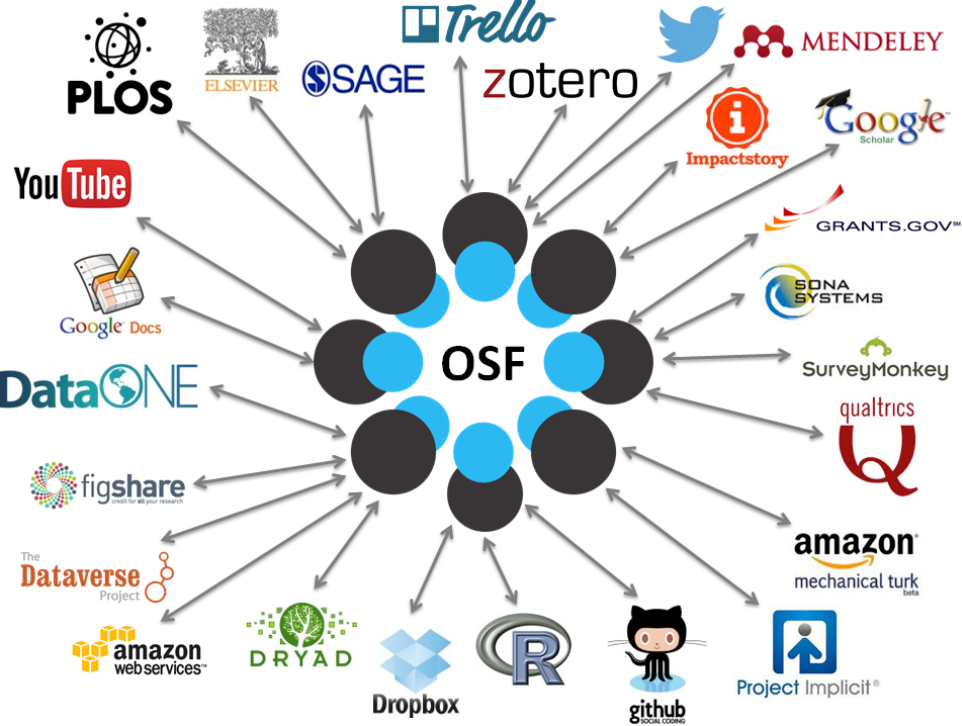
- Open Science Framework <http://osf.io>

- Open format

- Will soon sync with above

- Version Control!







- ◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ▶ ↺ 🔍 ↻



P-Hacking

BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Implementation
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

Do people actually do this? (John, Loewenstein, Prelec
2011)

1. In a paper, failing to report all of a study's dependent measures
2. Deciding whether to collect more data after looking to see whether the results were significant
3. In a paper, failing to report all of a study's conditions
4. Stopping collecting data earlier than planned because one found the result that one had been looking for
5. In a paper, "rounding off" a p value (e.g., reporting that a p value of .054 is less than .05)
6. In a paper, selectively reporting studies that "worked"
7. Deciding whether to exclude data after looking at the impact of doing so on the results
8. In a paper, reporting an unexpected finding as having been predicted from the start
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)
10. Falsifying data

1. In a paper, failing to report all of a study's dependent measures	63.4
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9
3. In a paper, failing to report all of a study's conditions	27.7
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6
5. In a paper, "rounding off" a p value (e.g., reporting that a p value of .054 is less than .05)	22.0
6. In a paper, selectively reporting studies that "worked"	45.8
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0
10. Falsifying data	0.6

	Admission rate	Defensibility rate
1. In a paper, failing to report all of a study's dependent measures	63.4	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	1.79 (0.44)
3. In a paper, failing to report all of a study's conditions	27.7	1.77 (0.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	1.76 (0.48)
5. In a paper, "rounding off" a p value (e.g., reporting that a p value of .054 is less than .05)	22.0	1.68 (0.57)
6. In a paper, selectively reporting studies that "worked"	45.8	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	1.61 (0.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	1.50 (0.60)
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	1.32 (0.60)
10. Falsifying data	0.6	0.16 (0.38)

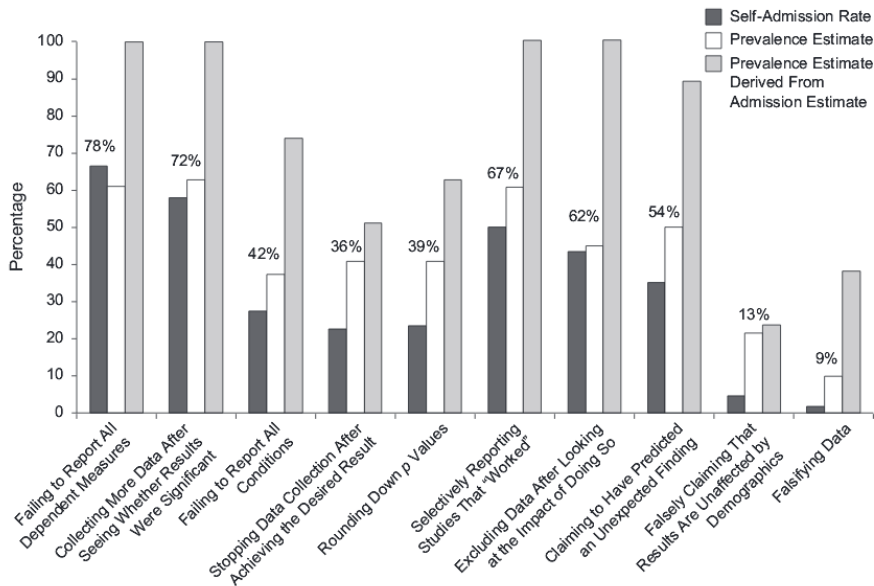


Fig. 1. Results of the Bayesian-truth-serum condition in the main study. For each of the 10 items, the graph shows the self-admission rate, prevalence estimate, prevalence estimate derived from the admission estimate (i.e., self-admission rate/admission estimate), and geometric mean of these three percentages (numbers above the bars). See Table 1 for the complete text of the items.

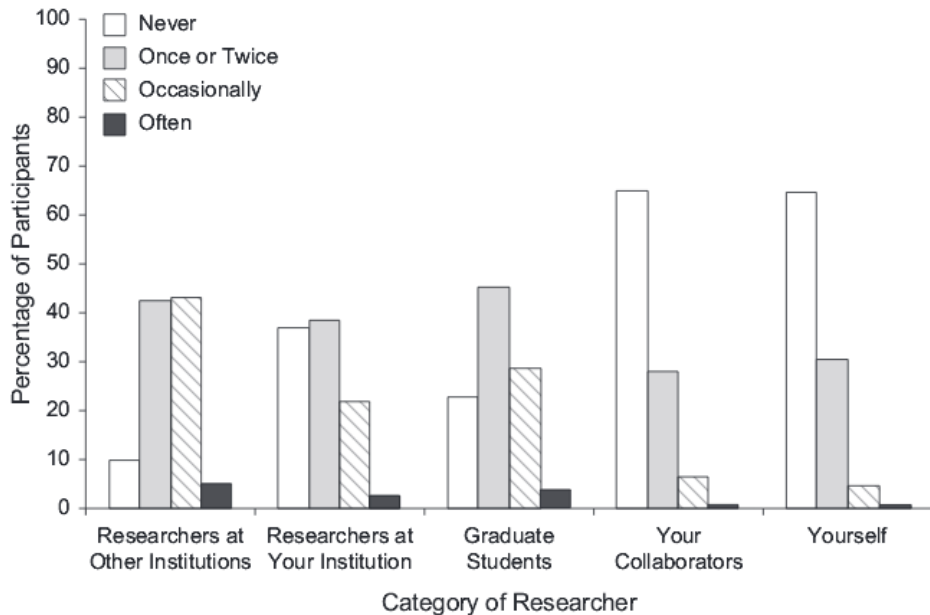


Fig. 2. Results of the main study: distribution of responses to a question asking about doubts concerning the integrity of the research conducted by various categories of researchers.



- “Science isn’t Broken” —538 journalism piece with interactive demo [Link](#)
- Listening to the Beatles’ “When I’m Sixty-Four” makes you younger. (Simmons, Nelson, Simonsohn 2011)
- Inordinately many .049 p-values, and indordinately few .051’s; 10-20%. (Brodeur et al 2015, “Star Wars”)
- Political ideologues literally see in black and white (Nosek, Spies, Motyl 2012)



Pre-Analysis Plan

Implementation Registration and Pre-Analysis Plans

Garret Christensen

Outline

Pre-Analysis Plan

Conclusion

- Often part of a registration
- From 3ie: “A pre-analysis plan is a detailed description of the analysis to be conducted that is written in advance of seeing the data on impacts of the program being evaluated. It may specify hypotheses to be tested, variable construction, equations to be estimated, controls to be used, and other aspects of the analysis. A key function of the pre-analysis plan is to increase transparency in the research. By setting out the details in advance of what will be done and before knowing the results, the plan guards against data mining and specification searching. Researchers are encouraged to develop and upload such a plan with their study registration, but it is not required for registration.”



BITSS

BIOMEDICAL INSTITUTE FOR TRANSDISCIPLINARY
SCIENCE

Origin: FDA's Guidance for Industry

Implementation
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

“E9 Statistical Principles for Clinical Trials” (1998) [▶ Link](#) §V Data Analysis Considerations

- 1 Prespecification of the Analysis
- 2 Analysis Sets
- 3 Missing Values and Outliers
- 4 Data Transformation
- 5 Estimation, Confidence Intervals, and Hypothesis Testing
- 6 Adjustment of Significance and Confidence Levels
- 7 Subgroups, Interactions, and Covariates
- 8 Integrity of Data and Computer Software Validity



BITSS

BIOMEDICAL INSTITUTE FOR TRANS-PREVENTION
IN THE SOCIAL SCIENCES

Glennerster, Takavarasha Suggestions

Implementation
Registration
and
Pre-Analysis
Plans

Garret
Christensen

Outline

Publication
Bias

Registrations

P-Hacking

Pre-Analysis
Plan

Conclusion

Running Randomized Evaluations

- 1 the main outcome measures,
- 2 which outcome measures are primary and which are secondary,
- 3 the precise composition of any families that will be used for mean effects analysis,
- 4 the subgroups that will be analyzed,
- 5 the direction of expected impact if we want to use a one-sided test, and
- 6 the primary specification to be used for the analysis.

World Bank Development Impact Blog

- 1 Description of the sample to be used in the study
- 2 Key data sources
- 3 Hypotheses to be tested throughout the causal chain
- 4 Specify how variables will be constructed
- 5 Specify the treatment effect equation to be estimated
- 6 What is the plan for how to deal with multiple outcomes and multiple hypothesis testing?
- 7 Procedures to be used for addressing survey attrition
- 8 How will the study deal with outcomes with limited variation?
- 9 If you are going to be testing a model, include the model
- 10 Remember to archive it

Wide range of when to write and how detailed to make the plan. At the extreme level of detail you would have your entire code already written before you got any data.

- J-PAL Hypothesis Registry (11), see <http://www.povertyactionlab.org/Hypothesis-Registry>
- AEA Registry has relatively few, plentiful in EGAP.
- Casey, Glennerster, Miguel, “Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan” *QJE* 2012. (Paper, Plan)
 - Government-sponsored program.
 - Broad program (Community Driven Development)
 - Broad outcomes (trust, public goods, public services, community groups, information, participation, crime, welfare, attitudes)

Outcome variable	(1) Mean for controls	(2) Treatment effect
Panel A: GoBifo “weakened” institutions		
Attended meeting to decide what to do with the tarp	0.81	−0.04 ⁺
Everybody had equal say in deciding how to use the tarp	0.51	−0.11 ⁺
Community used the tarp (verified by physical assessment)	0.90	−0.08 ⁺
Community can show research team the tarp	0.84	−0.12 [*]
Respondent would like to be a member of the VDC	0.36	−0.04 [*]
Respondent voted in the local government election (2008)	0.85	−0.04 [*]
Panel B: GoBifo “strengthened” institutions		
Community teachers have been trained	0.47	0.12 ⁺
Respondent is a member of a women’s group	0.24	0.06 ^{**}
Someone took minutes at the most recent community meeting	0.30	0.14 [*]
Building materials stored in a public place when not in use	0.13	0.25 [*]
Chiefdom official did not have the most influence over tarp use	0.54	0.06 [*]
Respondent agrees with “Responsible young people can be good leaders” and not “Only older people are mature enough to be leaders”	0.76	0.04 [*]
Correctly able to name the year of the next general elections	0.19	0.04 [*]



Why?

Garret Christensen

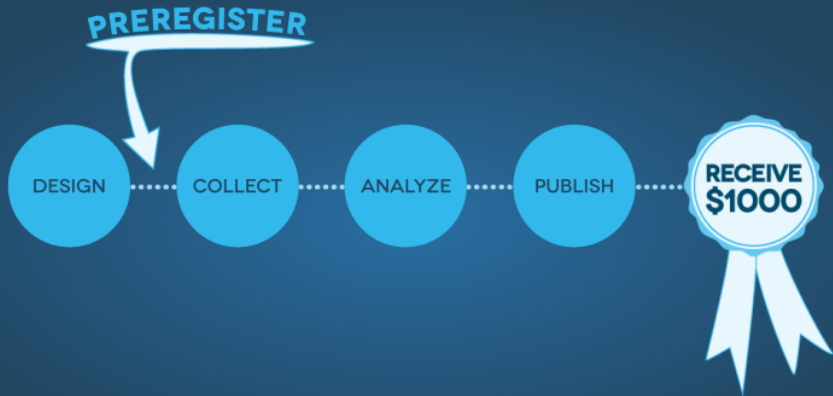
P-Hacking

Conclusion

- It's good science; you can distinguish between confirmatory and exploratory analysis.
- You get a badge.
 - Project Page. [Link](#)
 - Fertile women aren't more racist (Hawkins, Fitzgerald, Nosek 2015).
- You get \$1000. [Link](#)



CENTER FOR OPEN SCIENCE



THE PREREGISTRATION CHALLENGE

Learn more at cos.io/prereg

- Register your work to reduce publication bias.
- Include a pre-analysis plan to reduce researcher degrees of freedom.
- Register in most appropriate site for your work.
 - AEA
 - EGAP
 - RIDIE
 - OSF will hold anything, and link to your entire workflow.