Open science

Macartan Humphreys

Section 1

Experimentation: Processes and Workflows

Experimentation: Processes and Workflows

- Scope for experimentation
- Ethics of experiments
- Open science workflows

Subsection 1

When to experiment

Prospects

- Whenever someone is uncertain about *something* they are doing (all the time)
- Whenever someone hits scarcity constraints
- When people have incentives to demonstrate that they are doing the right thing (careful...)

Prospects

- **Advice**: If you can, **start from theory** and find an intervention, rather than the other way around.
- **Advice**: If you can, go for *structure* rather than *gimmicks*
- **Advice**: In attempts to parse, beware of generating unnatural interventions (how should a voter think of a politician that describes his policy towards Korea in detail but does not mention the economy? Is not mentioning the economy sending an unintended message?)

Prospects & Potential

- Randomization of where police are stationed (India)
- Randomization of how government tax collectors get paid (do they get a share?) (Pakistan)
- Randomization of the voting rules for determining how decisions get made (Afghanistan)
- Random assignment of populations to peacekeepers (Liberia)
- Random assignment of ex-combatants out of their networks (Indonesia)
- Randomization of students to ethnically homogeneous or ethnically diverse schools (anywhere?)

Subsection 2

Ethics

Constraint: Is it ethical to manipulate subjects for research purposes?

- There is no foundationless answer to this question. So let's take some foundations from the Belmont report and seek to ensure:
  1. Respect for persons
  2. Beneficence
  3. Justice
- Unfortunately, operationalizing these requires further ethical theories. Let's assume that (1) is operationalized by informed consent (a very liberal idea). We are a bit at sea for (2) and (3) (the Belmont report suggests something like a utilitarian solution).
- The major focus on (1) by IRBs might follow from the view that if subjects consent, then they endorse the ethical calculations made for 2 and 3 — *they* think that it is good and fair.
- This is a little tricky, though, since the study may not be good or fair

Is it ethical to manipulate subjects for research purposes?

- The problem is that many (many) field experiments have nothing like informed consent.
- For example, whether the government builds a school in your village, whether an ad appears on your favorite radio show, and so on.
- Consider three cases:
  1. You work with a nonprofit to post (true?) posters about the crimes of politicians on billboards to see effects on **voters**
  2. You hire confederates to offer bribes to **police officers** to see if they are more likely to bend the law for coethnics
  3. The British government asks you to work on figuring out how the use of water cannons helps stop **rioters** rioting

Is it ethical to manipulate subjects for research purposes?

- Consider three cases:
  - You work with a nonprofit to post (true?) posters about the crimes of politicians on billboards to see effects on **voters**
  - You hire confederates to offer bribes to **police officers** to see if they are more likely to bend the law for coethnics
  - The British government asks you to work on figuring out how the use of water cannons helps stop **rioters** rioting
- In all cases, there is **no consent** given by subjects.
- In 2 and 3, the treatment is **possibly harmful** for subjects, and the results might also be harmful. But even in case 1, there could be major unintended harmful consequences.
- In cases 1 and 3, however, the "intervention" is within the sphere of **normal activities** for the implementer.

Constraint: Is it ethical to manipulate subjects for research purposes?

- Sometimes it is possible to use this point of difference to make a "spheres of ethics" argument for "embedded experimentation."
- **Spheres of Ethics Argument**: Experimental research that involves manipulations that are not normally appropriate for researchers may nevertheless be ethical if:
  - Researchers and implementers agree on a **division of responsibility** where implementers take on responsibility for actions
  - Implementers have **legitimacy** to make these decisions within the sphere of the intervention
  - Implementers are indeed **materially independent** of researchers (no swapping hats)
- Difficulty with this argument:
  - **Question begging**: How to determine the legitimacy of the implementer? (Can we rule out Nazi doctors?)

Subsection 3

Transparency & Experimentation

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Analytic replication.** This should be a no brainer. Set everything up so that replication is easy. Use rmarkdown, or knitr or sweave. Or produce your replication code as a package.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Data.** How soon should you make your data available? **My view**: as soon as possibe. Along with working papers and before publication. Before it affects policy in any case. Own the ideas not the data.
    - Hard core: no citation without (analytic) replication. Perhaps. Non-replicable results should not be influencing policy.
- **Where should you make your data available?** Dataverse is focal for political science. Not personal website (mea culpa)
- **What data should you make available?** Disagreement is over how raw your data should be. **My view**: as raw as you can but at least post cleaning and pre-manipulation.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Should you register?**: Hard to find reasons against. But case strongest in testing phase rather than exploratory phase.
- **Registration**: When should you register? **My view:** Before treatment assignment. (Not just before analysis, mea culpa)
- **Registration**: Should you deviate from an preanalysis plan if you change your mind about optimal estimation strategies. **My view:** Yes, but make the case and describe both sets of results.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Registration**: When should you register? **My view:** Before treatment assignment. (Not just before analysis, mea culpa)
- **Registration**: Should you deviate from an preanalysis plan if you change your mind about optimal estimation strategies. **My view:** Yes, but make the case and describe both sets of results.

Subsection 4

Pre-registration rationales and structures

Two distinct rationales for registration

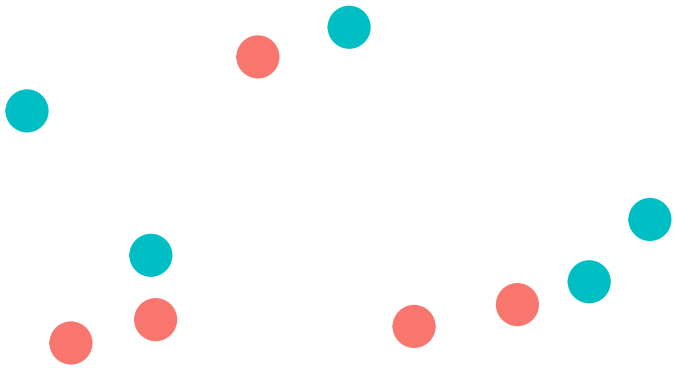
- File drawer bias (Publication bias)
- Analysis bias (Fishing)

File drawer bias

– Say in truth X affects Y in 50% of cases.

– Researchers conduct multiple excellent studies. But they only write up the 50% that produce "positive" results.

– Even if each individual study is indisputably correct, the account in the research record – that X affects Y in 100% of cases – will be wrong.

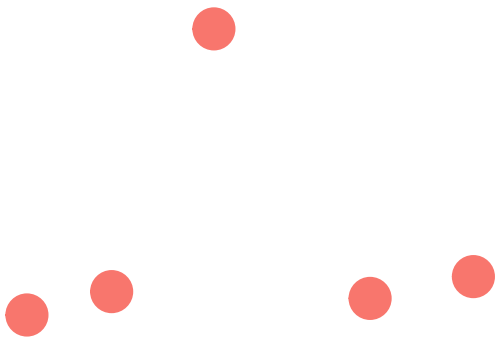


File drawer bias

– Say in truth X affects Y in 50% of cases.

– Researchers conduct multiple excellent studies. But they only write up the 50% that produce "positive" results.

– Even if each individual study is indisputably correct, the account in the research record – that X affects Y in 100% of cases – will be wrong.



File drawer bias

Exacerbated by:

– Publication bias – the positive results get published

– Citation bias – the positive results get read and cited

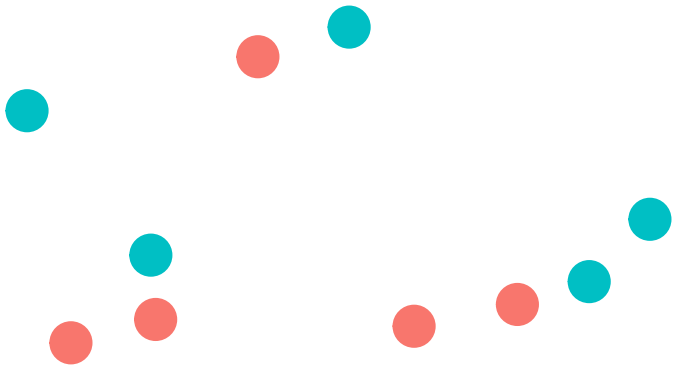
– Chatter bias – the positive results gets blogged, tweeted and TEDed.

Analysis bias (Fishing)

– Say in truth X affects Y in 50% of cases.

– But say that researchers enjoy discretion to select measures for X or Y, or enjoy discretion to select statistical models after seeing X and Y in each case.

– Then, with enough discretion, 100% of analyses may report positive effects, even if all studies get published.

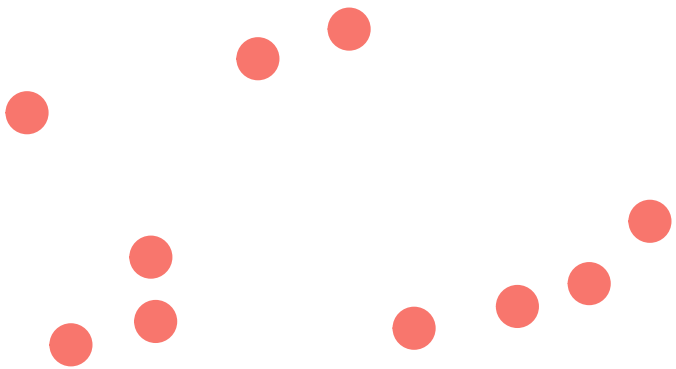


Analysis bias (Fishing)

– Say in truth X affects Y in 50% of cases.

– But say that researchers enjoy discretion to select measures for X or Y, or enjoy discretion to select statistical models after seeing X and Y in each case.

– Then, with enough discretion, 100% of analyses may report positive effects, even if all studies get published.



Analysis bias (Fishing)

– Try the exact fishy test An Exact Fishy Test (https://macartan.shinyapps.io/fish/)

– What's the problem with this test?

Evidence-Proofing: Illustration

- When your conclusions do not really depend on the data
- Eg – some evidence will always support your proposition – some interpretation of evidence will always support your proposition
- Knowing the mapping from data to inference in advance gives a handle on the false positive rate.

Evidence Proofing: Bayesian Illustration

- Say choice of two pieces of evidence to bring to bear, $K1$ or $K2$

Table 1: Likelihoods

(a)

(b) If TRUE

	K_1 = No	K_1 = Yes	All
K_2 = No	0.9	0.05	0.95
K_2 = Yes	0.05	0	0.05
All	0.95	0.05	1

(c)

(d) If FALSE

	K_1 = No	K_1 = Yes	All
K_2 = No	0.9	0.05	0.95
K_2 = Yes	0.05	0	0.05
All	0.95	0.05	1

- Posterior | K1 = Posterior | K2 = 95%
- Probability positive claim | H is false; evidence randomly selected (p) = 5%
- Probability positive claim | H is false; evidence is fished (p) = 10%

Evidence Proofing: Bayesian Illustration

- Say choice of two pieces of evidence to bring to bear, $K1$ or $K2$

Table 2: Likelihoods

(a)

(b) If TRUE

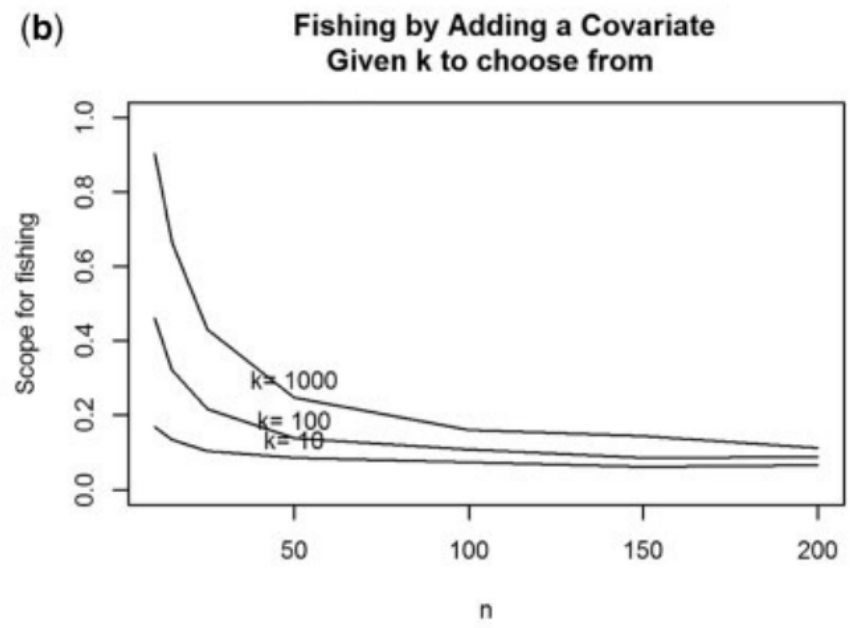
(c)

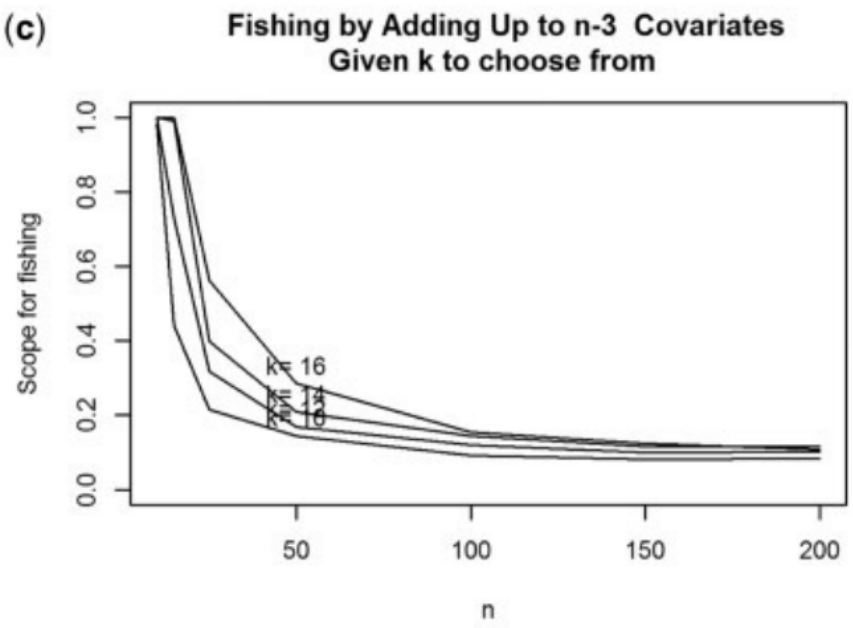
(d) If FALSE

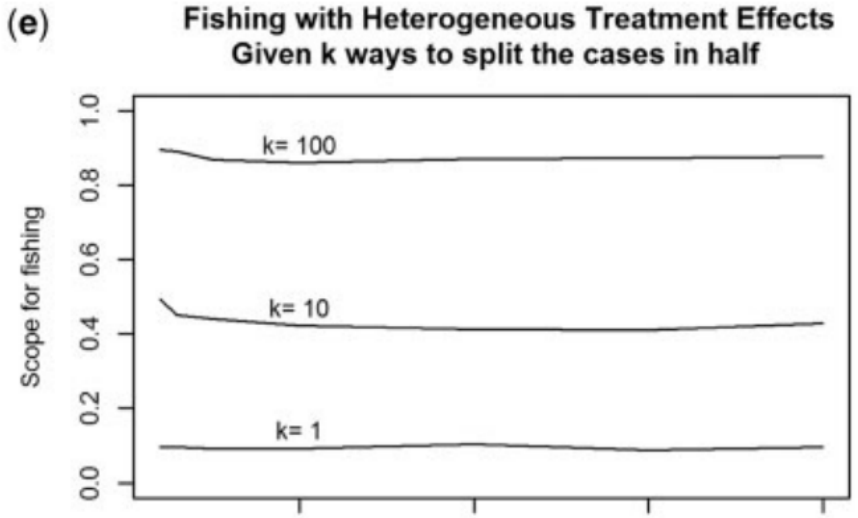
	K_1 = No	K_1 = Yes	All	K_1 = No	K_1 = Yes
K_2 = No	0.9	0.05	0.95	0	
K_2 = Yes	0.05	0	0.05	0.05	
All	0.95	0.05	1	0.05	

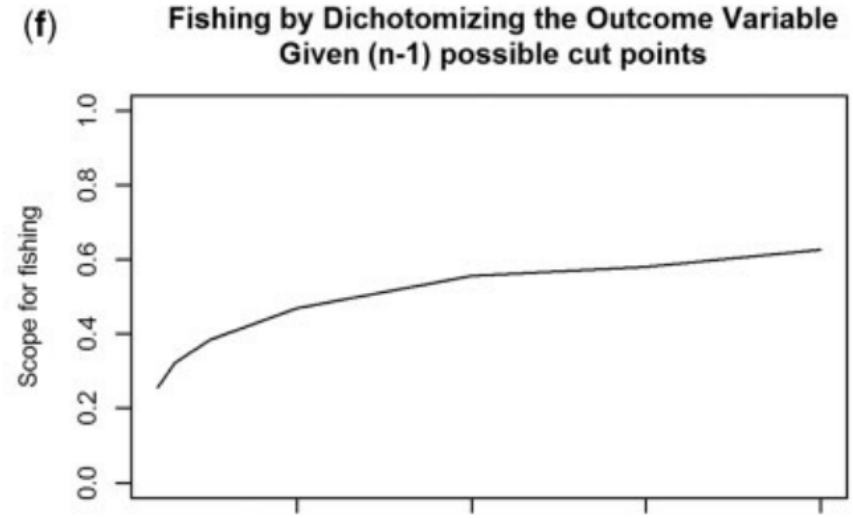
- What's the truly correct inference if you KNOW that researcher is a fisher?
- Depends: say you thought K1 and K2 were sought in order. Then if K2 evidence is presented this means K1 not found. So posterior |K2

The scope for fishing

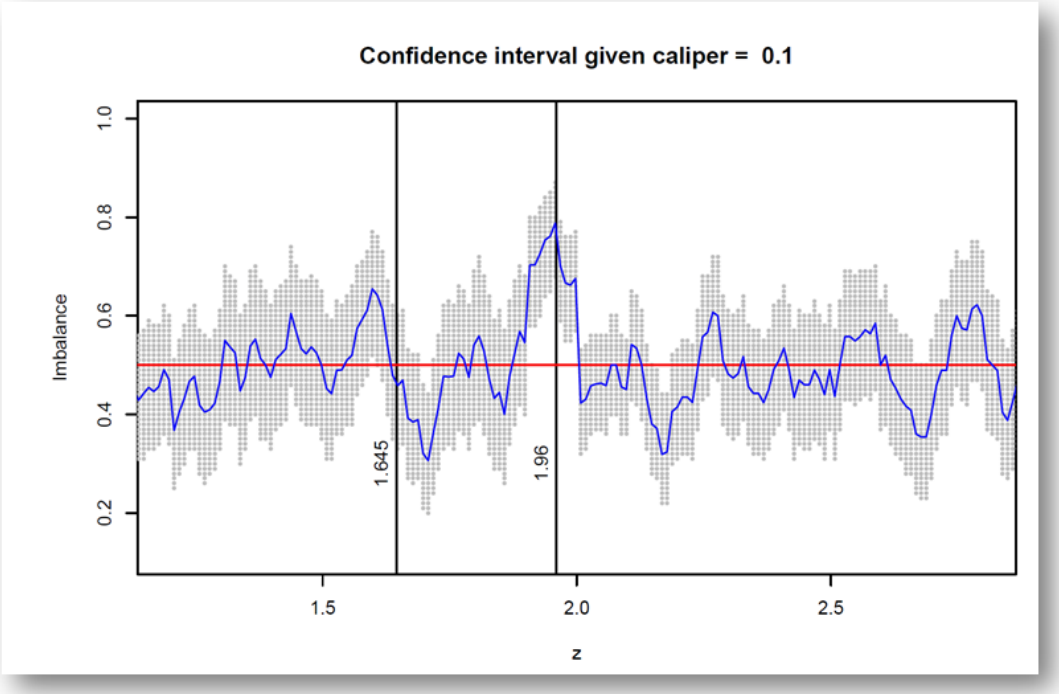








Evidence from political science



Source: Gerber and Malhotra

More evidence from TESS

- Malhotra tracked 221 TESS studies.
- 20% of the null studies were published. 65% not even written up (file drawer or anticipation of publication bias)
- 60% of studies with strong results were published.

Implications are:

- population of results not representative
- (subtler) individual published studies are also more likely to be overestimates

The problem

- Summary: we do not know when we can or cannot trust claims made by researchers.
- [Not a tradition specific claim]

Registration as a possible solution

Simple idea:

– It's about communication: – just say what you are planning on doing before you do it – if you don't have a plan, say that – If you do things differently from what you were planning to do, say that

Bells and whistles

– To be really useful a registry would have to have some credibility, some searchability, and some consistency in fields.

Registration as a possible solution

Elements:

- Make it a facility
- Non-mandatory
- Non-binding
- But comprehensive
- Report whether registered or not
- Report changes in plans

What's the Right Scope:

For discussion: but claims of "tests" seem like a good start

Bells and Whistles: Certificaton?

Center for Open Science Badges

- A public date-time stamped registration is in an institutional registration system (e.g., ClinicalTrials.gov, Open Science Framework)
- Registration pre-dates realization of the outcomes
- Registered design and analysis plan corresponds directly to reported design and analysis
- Full disclosure of results following the registered plan

Notations PR (peer review certified), DE (data exist), and TC (transparent changes)

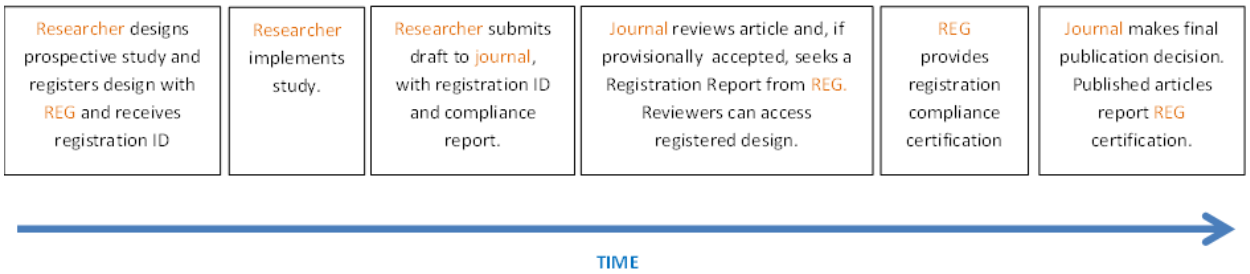
- https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/

Bells and Whistles: Certificaton?

Center for Open Science Badges

- Requires validating body (eg journal or registry)
- Validation on process not on quality
- Emphasis on transparency

Possible Cycle



Possible Models

- Journal-led model: let the registry follow

– Hard form – Medical sciences for RCTs – Soft form – Medical sciences for observational studies

- Professional Association led model

– AEA for RCTs – APSA?

- Funder led model (more mandatory) – RIDIE, NSF?
- Bottom up model? – Eg established by APSA sections; CQRM, PolMeth, Experiments? No formal journal recognition.

But even the simple idea is not everywhere welcome. There are many worries and some myths.

Subsection 5

Worries and Myths around registration

Myth: Concerns about fishing presuppose researcher dishonesty

- Fishing can happen in very subtle ways, and may seem natural and justifiable.
- Example:

– I am interested in whether more democratic institutions result in better educational outcomes. – I examine the relationship between institutions and literacy and between institutions and school attendance. – The attendance measure is significant and the literacy one is not. Puzzled, I look more carefully at the literacy measure and see various outliers and indications of measurement error. As I think more I realize too that literacy is a slow moving variable and may not be the best measure anyhow. I move forward and start to analyze the attendance measure only, perhaps conducting new tests, albeit with the same data.

Structural challenge

Our journal review process is largely organized around advising researchers how to adjust analysis in light of findings in the data.

Myth: Fishing is technique specific

- Frequentists can do it
- Bayesians can do it too.
- Qualitative researchers can also do it.
- You can even do it with descriptive statistics

Myth: Fishing is estimand specific

- You can do it when estimating causal effects
- You can do it when studying mechanisms
- You can do it when estimating counts

Myth: Registration only makes sense for experimental studies, not for observational studies

- The key distinction is between prospective and retrospective studies.
- Not between experimental and observational studies.
- A reason (from the medical literature) why registration is especially important for experiments: because you owe it to subjects
- A reason why registration is less important for experiments: because it is more likely that the intended analysis is implied by the design in an experimental study. Researcher degrees of freedom may be greatest for observational qualitative analyses.

Worry: Registration will create administrative burdens for researchers, reviewers, and journals

- Registration will produce some burden but does not require the creation of content that is not needed anyway

– It does shift preparation of analyses forward – And it also can increase the burden of developing analyses plans even for projects that don't work. But that is in part, the point.

- Upside is that ultimate analyses may be much easier.

Worry: Registration will force people to implement analyses that they know are wrong

- Most arguments for registration in social science advocate for non-binding registration, where deviations from designs are possible, though they should be described.
- Even if it does not prevent them, a merit of registration is that it makes deviations visible.

Myth: Replication (or other transparency practices) obviates the need for registration

- There are lots of good things to do, including replication.
- Many of these do not substitute for each other. (How to interpret a fished replication of a fished analysis?)
- And they may likely act as complements
- Registration can clarify details of design and analysis and ensure early preparation of material. Indeed material needed for replication may be available even before data collection

Worry: Registration will put researchers at risk of scooping

- But existing registries allow people to protect registered designs for some period
- Registration may let researchers lay claim to a design

Worry: Registration will kill creativity

- This is an empirical question. However, under a nonmandatory system researchers could:
- Register a plan for structured exploratory analysis
- Decide that exploration is at a sufficiently early stage that no substantive registration is possible and proceed without registration.

Implications:

- In neither case would the creation of a registration facility prevent exploration.
- What it might do is make it less credible for someone to claim that they have tested a proposition when in fact the proposition was developed using the data used to test it.
- Registration communicates when researchers are angage in exploration or not. We love exploration and should be proud of it.

The challenge of historical data

- Does registering analyses of historical data make sense?
- The problem is not just that researchers might have already seen the testing data; but that they have seen data that is correlated with it.

Historical data: Illustration

- Consider historical proposition H. – Say we start with a prior of .5 that H is true. – Say that if H is true then we observe K1 with probability 0.8 but if it is false we observe K1 with probability 0.2 ("double decisiveness") – Similarly if H is true then we observe K2 with probability 0.8 but if it is false we observe K2 with probability 0.2 ("double decisiveness" again)
- Say we observe K1 (some collection of facts)
- We then update our belief in H...

Historical data: Illustration

- Our updated belief is:

$$\Pr(H|K1) = \Pr(K1|H)\Pr(H)/\Pr(K1) = \frac{.8 * .5}{.8 * .5 + .2 * .5} = 80\%$$

- We are now 80% confident in proposition H.
- We decide to look for evidence K2. And we find it!
- Our posterior is now:

$$\Pr(H|K2) = \Pr(K2|H)\Pr(H)/\Pr(K2) = .8*.8/(.8*.8+.2*.2) = 94\%$$

- Or is it?

Historical data: Illustration

- Our updated belief is:

$$\Pr(H|K1) = \Pr(K1|H)\Pr(H)/\Pr(K1) = .8*.5/(.8*.5+.2*.5) = 80\%$$

- We are now 80% confident in proposition H.
- We decide to look for evidence K2. And we find it!
- Our posterior is now:

$$Pr(H|K2) = Pr(K2|H)Pr(H)/Pr(K2) = .8*.8/(.8*.8+.2*.2) = 94$$

- Or is it?

Historical data: Illustration

- What if there are correlated probabilities?
- Then

$$\Pr(H|K1\&K2) = .76 \times .5/(.76 \times .5 + .16 \times .5) = 83\%$$

Historical data: Illustration

- In a sense the fishing has already happened.
- How so?
- Say the proposition is FALSE but K1 is still observed
- A decision is then made to seek "new data" K2
- Now K2 will be observed with 80% probability even though H is false

Historical data: Illustration

- Naïve inference (using a prior of 80% due to K1): 94% if K2; 50% if not K2
- Inference if K1 used to decide on search for K2 but prior is "reset" to .5 80% if K2; 20% if not K2
- Sophisticated inference: 83% if K2; 50% if not K2
- This sophisticated inference is unchanged if you take explicit account of the fact that searching for K2 was conditional on K1; either way it is still $\Pr(H|K1, K2)$.
- *It requires assessing the probability of knowing what you know now and finding out what you will find, if the proposition is true or false.*

Historical data: Illustration

- Naïve inference (using a prior of 80% due to K1): 94% if K2; 50% if not K2
- Inference if K1 used to decide on search for K2 but prior is "reset" to .5 80% if K2; 20% if not K2
- Sophisticated inference: 83% if K2; 50% if not K2
- This sophisticated inference is unchanged if you take explicit account of the fact that searching for K2 was conditional on K1; either way it is still Pr(H | K1, K2).
- *Can such beliefs be elicited?* Perhaps.

Will it make a difference?

- Striking paucity of evidence.

How to?

- Let's look at an example
- Design declaration idea

Subsection 6

Reconciliation

Reconciliation

Incentives and strategies

Reconciliation

Table 22.1: Illustration of an inquiry reconciliation table.

Inquiry	In the preanalysis plan	In the paper	In the appendix
Gender effect	X	X	
Age effect			X

Table 22.2: Illustration of an answer strategy reconciliation table.

Inquiry	Following *A* from the PAP	Following *A* from the paper	Notes
Gender effect	estimate = 0.6, s.e = 0.31	estimate = 0.6, s.e = 0.25	Difference due to change in control variables [provide cross references to tables and code]

Reconciliation

Subsection 7

Replication files