Pre-Analysis Plans, Registries, Power Calculations and Preparing to do Research

Jacob Ricker-Gilbert, PhD

Associate Professor
Dept. of Agricultural Economics
Purdue University
USA

MwAPATA seminar, 10 December 2020.

I. Overview: We can think of this talk as advice on ways to plan our research.

- Before we go to the field, before we begin our analysis, what are some steps we can take to make our research better?
- How can we collect data more efficiently?
    - eg: don't ask survey questions that will never be analyzed.
- How can we conduct research (collect data, do analysis) in a transparent and ethical
- How do we make sure we have enough statistical power to reject the null when it should be rejected?
- The topics discussed to day are not required but are best practices
    - Becoming required more and more frequently by journals and donors.
    - Likely will continue to be in the future.

Talk Outline

1. Over view/Introduction
2. Public research registries
3. Pre-analysis plans
4. Power Calculations

Pre-analysis plans: Relevant publications used for this talk come from Module 8.3 in Glennerster and Takavarasha (2013), with parts from Olken (2015), Duflo et al. (2020), and Janzen and Michler (2021).

First Issue: the "file drawer" problem?

- Research that finds statistically significant results more likely to get published than those that do not reject the null
  - Andrews and Kasy (2019) found that studies with results that are statistically significant at the 5% level are 30 times more likely to be published than results that are not significant.
- Research that finds no impact of an intervention goes in the "file drawer" and never gets published.
- Obvious to see what that does to people's incentives?
- Also problematic because drawing overall conclusions about a particular impact?
  - For example, what are we missing if we are reviewing papers on a topic and all of the published literature finds statistically significant impacts?
  - Is that because they all find statistically significant impacts or is it because only the studies with statistically significant impacts get published?
    - Think about land titling impacts on agricultural productivity in Africa.
- One important key to research is to figure out a way to write an interesting paper that finds results that are not statistically significant.

This “publication bias” towards significant results leads to next problem. (Glennerster & Takavarasha)

Recap on statistics

- If we find a result is significant at the 5% level, what does this mean?
  - there is a 5% or less probability it is the result of chance
- If we test 10 independent hypotheses there is a 40% chance we will fail to reject the null at the 5% level at least once
  - i.e. 40% chance find one hypothesis is significant at the 5% level

Basic fact of statistics, the more regressions we run, the more likely we are to pick up statistical significance.

What does this give us incentive to do?

- Keep running regressions or dividing data until we find something.
  - If enough evaluations are run on a given type of program, some will give positive result by chance
  - Those with a vested interest may even deliberately run many studies and publicize only some of the results
  - Published results will suggest the program is effective even if it isn’t

Second issue: What is the publication bias problem cause?

- People to have incentive to keep searching until they find a statistically significant result.

- May be spurious or “cherry picked”

- The “grey area” comes in with where to draw the line.

    - More on this in a minute.

Data mining:

- Looking at the data many different ways, trying to find the result you want
- If test impact of program on many different outcomes, some will show positive (or negative) impact by chance
- If test impact of program on many different subgroups, some will show positive (or negative) impact by chance
- We may be falsely accused of data mining
  - E.g. we test one subgroup and report the results but readers think maybe we tested many and only reported the one that was significant

So far we have identified two problems

1. File drawer / publication bias problem
   - Statistically significant results more likely to get published.
   - Insignificant results are results but aren't made public so full body of evidence not available

2. Cherry picking / Data mining problem
   - Ex post adjustment of the design and/or analysis to tell a compelling story.
   - Running many models until you find a result

What are the solutions?

I. Research Registry of a study.

- Deals with the “file drawer” problem #1
- Registering at AEA RCT registry or other registries prior to doing field work.
  - Mandatory fields in registry : title, country, status, keyword(s), abstract, trial and intervention start and end dates, primary outcomes, experimental design, randomization method, randomization unit, clustering, sample size, IRB information, etc
  - Submissions are time stamped.
  - If collecting primary data, important to do before going to the field
  - Not always possible if using secondary data

- While AEA registry designed for RCT, Janzen and Michler argue that all studies that analyze data and test hypotheses should register.
  - Record of what analysis is out there and what has been done.
    - Example, many studies in Africa use LSMS data to look at impacts of agricultural technology adoption.
      - Is useful to know what technologies have been found to have impacts and which have not.

- AEA Social Science RCT registry

  https://www.socialscienceregistry.org/

- Below are links to other registries that are not limited to RCT's
- Identified in Duflo et al. (2020) pg. 11

There are other social science registries that accept studies that are not RCTs:

- The International Initiative for Impact Evaluation's (3ie) Registry for International Development Impact Evaluations (RIDIE), launched in September 2013, accepts any type of impact evaluation (including non-experimental) related to development in low- and middle-income countries. (https://ridie.3ieimpact.org/).
- The Evidence in Governance and Politics (EGAP) registry was also created in 2013 by a network of researchers focusing on governance, politics, and institutions. It accepts non-experimental designs as well and does not apply restrictions on the type of research. (http://egap.org/content/registration).
- The Center for Open Science's (COS) Open Science Framework (OSF) accommodates the registration of essentially any study or research document by allowing users to create a time-stamped web URL. Registrations cannot be edited or deleted. (https://osf.io/).

- Pre-Registry Also serves as a concise record of initial intentions (Duflo et al. 2020)

- Many top econ journals require registry of trial before submission to journal.
  - Doesn't have to be registered before data collected to submit to journals.

- Duflo et al. and other have suggested that if you register, do your study and for whatever reason you choose not to publish, write a short post-trial detail to record your reasons for not publishing and mark the study complete.
  - Also if something goes wrong you can write withdrawn and explain why.
  - Helps researchers in the future avoid same pitfalls.

- These are ways to avoid the file draw problem.
- But doing this work is a bit of a public good so…….
  - Likely under-provided.

II. Second Problem

The use of Pre-Analysis Plans (PAP) included in the registry can help reduce the data mining “cherry picking” problem.

- Write down in advance how the data will be analyzed
  - What outcomes are of primary interest?
  - What subgroups will be examined?
- Include the PAP with the pre-registry.
- When presenting results in paper, show all those covered in the PAP
  - highlight any deviations from or additions to the PAP

When is a PAP most useful? (Glennerster & Takavarasha)

- When a study has a large number of outcomes with no obvious hierarchy of which are the most important
  - Think of subjective measures of welfare
- When researchers know they are interested in differential impact on different subgroups (heterogeneous treatment effects)
  - How do women, youth, poor benefit from an intervention?
- When researchers are concerned others will push them to find positive impacts
  - Think of an NGO or government entity wanting to find positive results of a policy or program.
- When want to adjust statistical tests for multiple hypothesis tests
  - Reviewers getting stricter about this now.

When should a PAP be written?

- Before the baseline? (if collecting primary data this is when Duflo et al. recommend)
  - Prevents us from learning about what questions have low response rates or inconsistent answers
- Before the intervention starts?
  - prevents researchers taking advantage of random shocks which happen to mainly impact treatment or comparison
  - Also prevents researchers thinking of new hypotheses, e.g. unthought-of of negative consequences
- Before looking at any data? Olken (2015) says this is when it must be done
  - Can be useful to look at comparison data to determine appropriate control variables
  - drop variables where little chance of improvement (e.g. 95% of control already do)

What should be covered in a PAP?

Checklist from Olken (2015)

Table 1

Pre-Analysis Plan Checklist

Item	*Brief description*
Primary outcome variable	The key variable of interest for the study. If multiple variables are to be examined, one should know how the multiple hypothesis testing will be done.
Secondary outcome variable(s)	Additional variables of interest to be examined.
Variable definitions	Precise variable definitions that specify how the raw data will be transformed into the actual variables to be used for analysis.
Inclusion/Exclusion rules	Rules for including or excluding observations, and procedures for dealing with missing data.
Statistical model specification	Specification of the precise statistical model to be used, hypothesis tests to be run.
Covariates	List of any covariates to be included in analysis.
Subgroup analysis	Description of any heterogeneity analysis to be performed on the data.
Other issues	Other issues include data monitoring plans, stopping rules, and interim looks at the data.

The focus on PAP's in the last 10 years has led to some very detailed PAPS that try to cover all contingencies and specifications.

PAP Basically becomes a paper without results

Journal of Development Economics will review PAPs without results to make an R&R or reject decision.

There are also drawbacks….

Disadvantages of a PAP (Glennerster & Takavarasha)

- Any analysis that is not included in the PAP has less credibility
  - Only do a PAP if you have the time to think it through carefully
- Sometimes patterns in the data tell consistent stories we never thought of
  - We might want the flexibility to pursue these
- With complex evaluations it can be hard to think through all outcome combinations and how analysis should proceed with each
  - We may do different secondary analysis if the impact is positive vs. negative
  - One option is to do PAP in stages, look at some data, then write another PAP

Duflo et al. (2020) argue for a simpler PAP

- Include what you would put in the RCT registry in the PAP
  - title, country, status, keyword(s), abstract, trial and intervention start and end dates, primary outcomes, experimental design, randomization method, randomization unit, clustering, sample size, IRB information, etc.
- Do not make the PAP a paper "without results"
  - These are distinct and should be treated as such
  - Can populate the PAP when have results in the registry to explain findings, but leave out details of the paper.
- Admit in PAP if you are uncertain about certain parts of the study
  - Sample size power calculations, measurement errors, response rates,
- They argue that deviations from the PAP in a paper are not a *prima facia* cause for concern.
  - Better than studies that do not have a PAP
  - Explain deviation from PAP in paper and reason for it.

Take home messages on pre-registry and PAPS

1. You do not have to register your study or write a PAP
2. But you will probably have a better study if you start working on it with the intention to do so.
   - Write out a PAP even if you don't register it.
3. Thinking about the research: (hypotheses, models to estimate, outcome variables, control variables, clustering strategy, sub-populations of interest, contingencies, and uncertainty to estimation) before collecting data, **you will be better off.**
   - Avoids forgetting to ask something in a survey
   - Avoids asking questions on a survey that are not used in an analysis
4. Helps you focus your analysis and writing results and conclusions to make life easier on the back end.
5. These issues are not going to go away so important to be aware of them.

III. Power calculations

Question: How much data do I need to collect?

Motivated by 2 factors:

1. The hypotheses that I want to test

2. Budget

Most of this comes from Bloom (1995) and Duflo, Glennerster and Kremer (2008)

PICS project

Question: Does adoption of improved storage technology affect quantity of grain that households store at harvest?

H_o: adoption of technology has no significant effect on quantity of grain stored by household.

Alternative is that it has an effect. Assume effect is positive.

- Want to make sure that we do not have a **type I error**:

  That null is rejected when it is true "false positive." Detecting an effect that is not present

- And a **type II error**:

  that the null is not rejected when it is false "false negative." Failing to reject an effect that is present.

Program Design I

- Randomly divide household up into treatment and control groups.
  - Treatment group gets bags
  - Control group does not.

$$Y_i = \alpha + \beta T + \epsilon_i \, .$$

Y = outcome for household i

β = difference in sample means from two groups (ATE estimate)

T = treatment (0 or 1)

ε = error term for household i

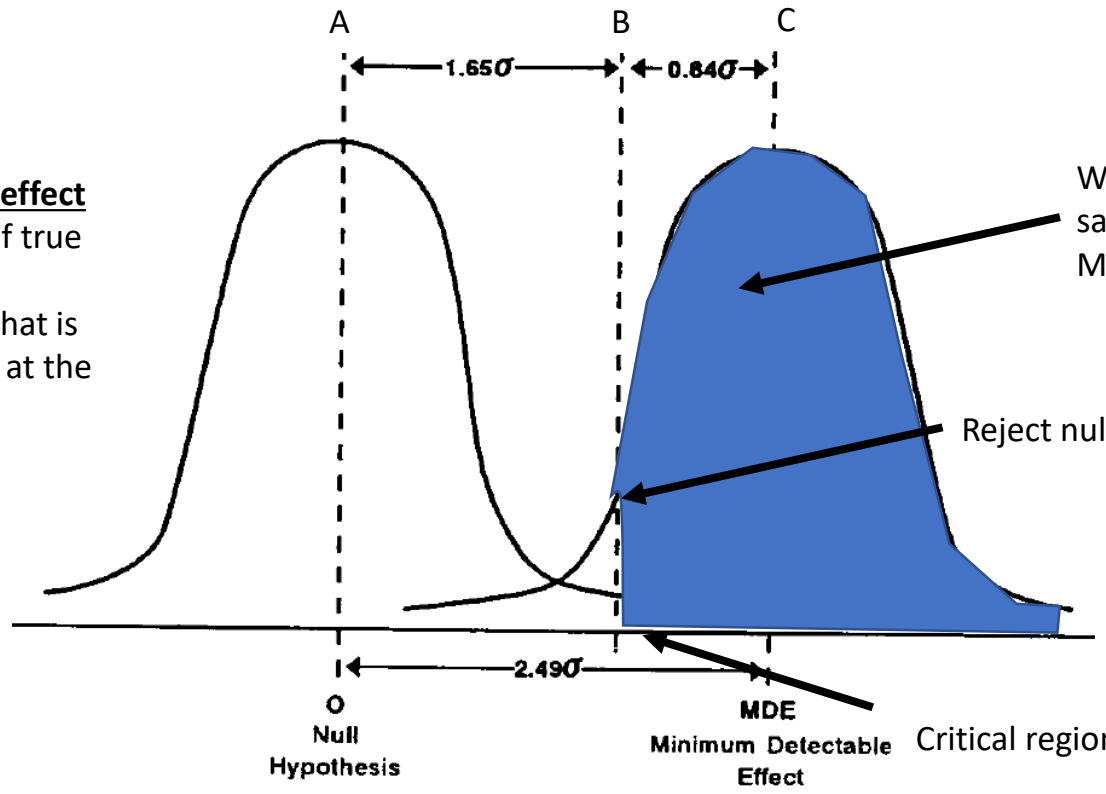
1 treatment and P proportion of the sample is treaded. Random sample from population so each obs. Is iid with variance σ^2

Variance of $\hat{\beta}$ = $\frac{1}{P(1-P)}\frac{\sigma^2}{N}$

Distribution of $\hat{\beta}$

- **0.05 for stat sign**
- **We want 80% power**

Minimum detectable effect = smallest effect that if true has an 80% chance of producing an impact that is statistically significant at the 5% level.



So MDE is 1.65+0.84 = 2.49 times the SE of the impact estimate

Example 1) from Bloom (2005)

Evaluating an experimental job training program

- SE of the impact estimate = $500
- MDE of the effect is 2.49 x $500 = $1,250
- Therefore, $1,250 is the true program impact with an 80% change of being identified (producing a significant and positive impact estimate at the 0.05 level)
- True positive impacts > than $1,250 have a > 80% chance of being identified
- True positive impacts < than $1,250 have a < 80% chance of being identified

Example 2) from Bloom (2005)

Evaluating an experimental job training program

<u>Design A</u>
MDE = 500

80% chance of picking up $500 increase as true impact

<u>Design B</u>
MDE = 1,000

80% chance of picking up $1,000 increase as true impact

Which design has more power?

To achieve power *k*:

$$\beta > (t_{1-\kappa} + t_{\alpha}) SE(\hat{\beta})$$

Minimum Detectable Effect size for a given power k is:

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}$$

α = significance level

N= sample size

P =proportion of subjects allocated to treatment group (P).

- Tradeoff between power and size. When N decreases then t_{α} increases.
- Tradeoff between falsely concluding that program has an effect when it doesn't, and probability of falsely concluding that it has no effect when it does.

Grouped Errors

- Often it is not practical to randomize interventions at the individual level.
- Easier to randomize at the group level (village, club, school, etc.)

- In this situation, people in the same group may be subject to a common shock, which means outcomes correlated.
  - If everyone in treatment village suffers drought it will not be possible to separate the drought effect from the program effect.

$$Y_{ij} = \alpha + \beta T + \upsilon_j + \omega_{ij}$$

For individual i in group j

j clusters of identical size n.

v_j is group error, iid with variance τ^2

w_{ij} is individual error, iid with variance σ^2

With Grouped randomization SE of $\hat{\beta}$ =

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2+\sigma^2}{nJ}}.$$

With individual randomization SE of $\hat{\beta}$ =

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}.$$

Design effect = $D = \sqrt{1+(n-1)\rho}$

$$\frac{\text{SE group level randomization}}{\text{SE Individual level randomization}}$$

Intracluster correlation = $\rho = \tau^2/(\tau^2+\sigma^2)$

Proportion of all variance explained by within group variance

Design effect increases with both intracluster correlation and number of people in each group.

Bloom (2005)

$$MDE = \frac{M_{J-2}}{\sqrt{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}}\,\sigma$$

$$M_{J-2} = t_{\alpha/2} + t_{1-\kappa},$$

- MDE affected by number of groups J
- If ρ is large then number of obs. per group matters much less.
- For given sample size, an increase in the number of individuals sampled per cluster increases precision much less than increasing the number of clusters.
- Total number of clusters to be sampled and number of people to sample per cluster are very dependent on ρ.

Command to compute Intracluster correlation (intraclass correlation) in Statais *icc*

Now Look at excel example.

Issues with Power Calculations

- Ultimately power calculations are only as accurate for your study as the data you use.
- If possible, good to do a pilot study in same area before larger study to quantify power needs.
  - Also understand the problem better.
- Being under-powered is a big problem that reviewers are keen on now days.
  - Thoughtful design important (clear objectives and hypotheses)
  - Collect baseline data. Get information on outcomes at baseline and include baseline outcomes as RHS variable
  - Stratify on areas where issues are more relevant (eg: poor villages, areas with higher aflatoxin levels, etc.)
  - More frequent follow-up to measure outcome variables at multiple points in time can increase power (McKenzie 2012, "The case for more T in experiments" JDE.)

Thank you for your time! Questions / Comments?

jrickerg@purdue.edu

References

Power Calculations

- Bloom, H. S. (1995): “Minimum detectable effects: A simple way to report the statistical power of experimental designs,” *Evaluation Review* 19:547–56.
- Duflo, E., R. Glennerster, and M. Kremer. 2008. “Using Randomization in Development Economics Research: A Toolkit.” *Handbook of Development Economics*. 4: 3,895-3,962.

Pre-Analysis Plans, Registeries

- Duflo, E., A. Banerjee, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautman (2020). “In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics.”Working Paper 26993, National Bureau of Economic Research.
- Glennerster, R. and K. Takavarasha (2013) Running Randomized Evaluations: A Practical Guidebook. Princeton University Press, Princeton, New Jersey.
- Janzen, S., and J. Michler (2021) “Ulysses' Pact or Ulysses' Raft: Using Pre-Analysis Plans in Experimental and Non-Experimental Research” (Forthcoming) *Applied Economics Prospective and Policy*
- Olken, B. A. (2015). “Promises and perils of pre-analysis plans.” *Journal of Economic Perspectives*. 29(3), 61-80.