Research Design, Randomization and Design-Based Inference

Paul Goldsmith-Pinkham

January 15, 2026

Outline on Randomization

- Discuss the value of randomized interventions, and identifying settings where interventions are “as-if” randomly assigned
  - Touch on the historical and (somewhat) current views on this
- Define a “research design.”
- Give an introduction to design-based vs. model-based identification and causal inference.

The power of randomization

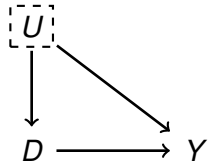
- Randomization is a powerful tool
  - E.g. An intervention giving a treatment to half of a sample using a randomized process
  - Formally, randomly assign D_i to a sample of size n such that the set of potential random assignments across all n individuals is known (Ω), and the probability distribution over Ω is known
  - In other words, you know the "true" propensity to receive treatment (the p-score)

- In our different models of causal inference:

The power of randomization

- Randomization is a powerful tool
  - E.g. An intervention giving a treatment to half of a sample using a randomized process
  - Formally, randomly assign D_i to a sample of size n such that the set of potential random assignments across all n individuals is known (Ω), and the probability distribution over Ω is known
  - In other words, you know the "true" propensity to receive treatment (the p-score)

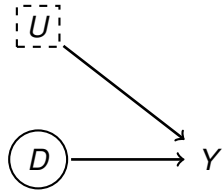
- In our different models of causal inference:
  - randomized intervention breaks paths on DAG



The power of randomization

- Randomization is a powerful tool
  - E.g. An intervention giving a treatment to half of a sample using a randomized process
  - Formally, randomly assign D_i to a sample of size n such that the set of potential random assignments across all n individuals is known (Ω), and the probability distribution over Ω is known
  - In other words, you know the "true" propensity to receive treatment (the p-score)

- In our different models of causal inference:
  - randomized intervention breaks paths on DAG



The power of randomization

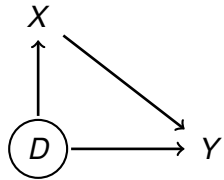
- Randomization is a powerful tool
  - E.g. An intervention giving a treatment to half of a sample using a randomized process
  - Formally, randomly assign D_i to a sample of size n such that the set of potential random assignments across all n individuals is known (Ω), and the probability distribution over Ω is known
  - In other words, you know the "true" propensity to receive treatment (the p-score)

- In our different models of causal inference:
  - randomized intervention breaks paths on DAG
  - Creates independence necessary for strong ignorability

The power of randomization

- Randomization is a powerful tool
  - E.g. An intervention giving a treatment to half of a sample using a randomized process
  - Formally, randomly assign D_i to a sample of size n such that the set of potential random assignments across all n individuals is known (Ω), and the probability distribution over Ω is known
  - In other words, you know the "true" propensity to receive treatment (the p-score)

- In our different models of causal inference:
  - randomized intervention breaks paths on DAG
  - Creates independence necessary for strong ignorability
  - Creates *some* forms of independence between the intervention and structural errors in a model
    - Why only some?



- Imagine an intervention that affects multiple outcomes
- Even randomized, if agents reoptimize with respect to X, this intervention no longer identifies the exclusive effect of D on Y without more assumptions

If the use of randomization is so powerful, why don't we always use it?

There are a few reasons:

1. People may not want to be randomized into different treatments.
   - They value their choices, and it may be impractical to randomize their decisions even if there is a clear benefit to doing so.
   - A firm, for example, may not want to randomize their policies
2. It may be unethical to randomize.
   - For example, if there is a clear benefit to a treatment, it may be unethical to withhold that treatment from individuals by placing them in the control.
3. It may be impossible to randomize.
   - For example, if we are interested in the effect of a policy change, it may be impossible to randomize the policy change across different regions or states.

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

> After three decades of churning out estimates, the econometrics club finds itself under critical scrutiny and faces incredulity as never before. Fischer Black writes of "The Trouble with Econometric Models." David Hendry queries "Econometrics: Alchemy or Science?" John W. Pratt and Robert Schlaifer question our understanding of "The Nature and Discovery of Structure." And Christopher Sims suggests blending "Macroeconomics and Reality."

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

> Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

> This image of the applied econometrician's art is grossly misleading. I would like to suggest a more accurate one. The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird droppings increase yields. However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing. The meeting reconvenes in the halls and in the bars, with heated discussion whether this is the kind of work that merits promotion from Associate to Full Farmer; the Luminists strongly opposed to promotion and the Aviophiles equally strong in favor.

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

> One should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data. The images I have drawn are deliberately prejudicial. First, we had the experimental scientist with hair neatly combed, wide eyes peering out of horn-rimmed glasses, a white coat, and an electronic calculator for generating the random assignment of fertilizer treatment to plots of land. This seems to contrast sharply with the nonexperimental farmer with overalls, unkempt hair, and bird droppings on his boots. Another image, drawn by Orcutt, is even more damaging: "Doing econometrics is like trying to learn the laws of electricity by playing the radio." However, we need not now submit to the tyranny of images, as many of us have in the past.

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

I. Is Randomization Essential?

What is the real difference between these two settings? Randomization seems to be the answer. In the experimental setting, the fertilizer treatment is "randomly" assigned to plots of land, whereas in the other case nature did the assignment. Now it is the tyranny of words that we must resist. "Random" does not mean adequately mixed in *every* sample. It only means that on the average, the fertilizer treatments are adequately mixed. Randomization implies that the least squares estimator is "unbiased," but that definitely does not mean that for each sample the estimate is correct. Sometimes the estimate is too high, sometimes too low. I am reminded of the lawyer who remarked that "when I was a young man I lost many cases that I should have won, but when I grew older I won many that I should have lost, so on the average justice was done."

A historical aside on the credibility revolution

- A director's cut of "Let's take the con out of econometrics" Leamer (1983)

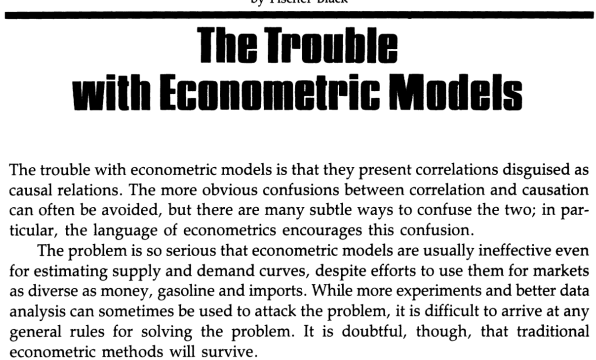
The truly sharp distinction between inference from experimental and inference from nonexperimental data is that experimental inference sensibly admits a conventional horizon in a critical dimension, namely the choice of explanatory variables. If fertilizer is randomly assigned to plots of land, it is conventional to restrict attention to the relationship between yield and fertilizer, and

to proceed as if the model were perfectly specified, which in my notation means that the misspecification matrix M is the zero matrix. There is only a small risk that when you present your findings, someone will object that fertilizer and light level are correlated, and there is an even smaller risk that the conventional zero value for M will lead to inappropriate inferences. In contrast, it would be foolhardy to adopt such a limited horizon with nonexperimental data. But if you decide to include light level in your horizon, then why not rainfall; and if rainfall, then why not temperature; and if temperature, then why not soil depth, and if soil depth, then why not the soil grade; ad infinitum. Though this list is never ending, it

A historical aside on the credibility revolution

- Important context for understanding current empirical methodology: empirics was viewed with tremendous skepticism by the 1980s

- Here's Black (1982)



by Fischer Black

The Trouble with Econometric Models

The trouble with econometric models is that they present correlations disguised as causal relations. The more obvious confusions between correlation and causation can often be avoided, but there are many subtle ways to confuse the two; in particular, the language of econometrics encourages this confusion.

The problem is so serious that econometric models are usually ineffective even for estimating supply and demand curves, despite efforts to use them for markets as diverse as money, gasoline and imports. While more experiments and better data analysis can sometimes be used to attack the problem, it is difficult to arrive at any general rules for solving the problem. It is doubtful, though, that traditional econometric methods will survive.

A historical aside on the credibility revolution

- Fast-forward 25 years later and Angrist and Pischke (2010) have declared a credibility revolution
- “Reseach design” is the clear victor, with pure randomization the leading champion

Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. Sensitivity analysis played a role in this, but as we see it, the primary engine driving improvement has been a focus on the quality of empirical research designs. This emphasis on research design is in the spirit of Leamer's critique, but it did not feature in his remedy.

The advantages of a good research design are perhaps most easily apparent in research using random assignment, which not coincidentally includes some of the most influential microeconometric studies to appear in recent years. For example,

What is a research design?

- A clear interpretation from this is that “research design” is important.
- Well, what’s the right definition for research design?
  - Shows up 69 times in Angrist and Pischke’s JEP piece, but not defined
- It seems almost “intuitive” but let’s try to define it.

David Card's definition of Research Design

- Card draws a distinction between causality as "model-based" and "design-based":
  - "causality is model-based: only exists within the framework of a theory that x causes y"
  - "causality is design-based: ...causality requires that you can design a manipulation in which x causes y"

- Crucial definition of what Card views as "design-based" approach:
  - "identification equated with research design"
  - "research design defines the counterfactual"
  - Of course, he also doesn't define (in his slides) what research design means...

- From Card's Nobel lecture: research design is equated with transparently describibng sources of identification

- `https://davidcard.berkeley.edu/lectures/woytinsky.pdf`

Paul Goldsmith-Pinkham's definition of Research Design

- A *(causal) research design* is a statistical and/or economic statement of how an empirical research paper will estimate a relationship between two (or more) variables that is causal in nature: how X causes Y.

- Since we know that causal effects require estimation of an (unobservable) counterfactual, a research design describes what assumptions are necessary to estimate the counterfactual(s) for a given estimand.

- As we will discuss in class, these research designs can be split into two types of assumptions (with some overlap to be discussed later):
  - *Model*-based: the estimand is identified using assumptions on the modeling of the potential outcomes conditional on treatment and additional variables (e.g. parallel trends)
  - *Design*-based: the estimand is identified using assumptions on the treatment variable, conditional on the potential outcomes and additional variables

- They are different *assumptions* to allow for credible estimates

Why was research design revolution so important?

- For today, we'll assume we have a randomized intervention: an example of a design-based approach
  - Ignore compliance
  - Ignore "quasi-experimental" vagaries
  - These are all solveable! See Bowers and Leavitt (2020) for discussion

- Knowledge of an explicit, randomized design provides a different approach to estimation and testing than what we traditionally learn in econometrics

- *Design-based* inference is
  1. Transparent
  2. Efficient

- Today: basic primer to give groundwork for rest of course
  - Very useful in some situations!

What is goal of design-based inference?

- Potential outcomes framework highlights that we can talk about every unit's PO.
  - Let there be a *finite* population of n individuals, $i = \{1, \ldots, n\}$
  - For each i, we have $(Y_i(0), Y_i(1), D_i)$, where $(Y_i(0), Y_i(1))$ denote their set of potential outcomes, and $D_i \in \{0, 1\}$ denote their treatment status
  - Let $\mathbf{Y}_0$ denote the vector of $Y_i(0)$, $\mathbf{Y}_1$ denote the vector of $Y_i(1)$, and $\mathbf{D}_0$ denote the vector of D_i.

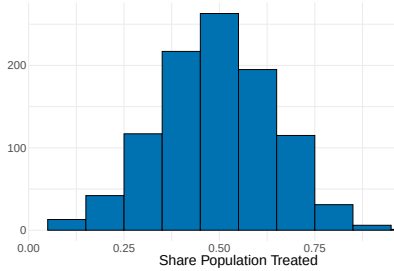
- What do we want to know / test about these outcomes?
  - Average? Distribution? Shifts? Underlying parameter?
  - For now, we'll focus on additive difference $\tau_i = Y_i(1) - Y_i(0)$, and the average of it $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau_i$.

- What do we want to do?
  - Let's start by making $\bar{\tau}$ our *estimand*

Define our research design

- Consider the set of potential ways that **D** could be randomized to the population
  - $\mathbf{Y}_1$ and $\mathbf{Y}_0$ are *fixed* – it is only the random variation in **D** that creates uncertainty

- Let Ω denote that space of possible values that **D** can take. It is defined by the type of randomize experiment one runs.
  - If each individual has a fair coin flipped on whether they are treatment or control, then $\Omega = \{0, 1\}^n$. But then the variation in number treated and control can vary quite a lot for small samples!
  - Other ways to consider randomly assigning individuals
    - Random draws from an urn (to ensure an exact number treated) – this is known as a *completely randomized experiment* – this is commonly used and convenient
    - Clustering individuals on characteristics (or location) – this is known as a *stratified randomized experiment*



Define our research design

- Key point: we know the exact probability distribution over Ω, and hence **D**.

Define our research design

- Key point: we know the exact probability distribution over Ω, and hence **D**.
- First consider with full knowledge for the true draw of **D** (the assignment that happened in our data)

D_i	$Y_i(1)$	$Y_i(0)$	Y_i
1	11.9	6.6	11.9
1	10	8.5	10
1	9.7	9.4	9.7
1	9.5	7	9.5
1	11.4	7.4	11.4
0	9.6	7.6	7.6
0	9.1	7.1	7.1
0	10.4	7.7	7.7
0	10.4	8	8
0	12.4	7.8	7.8

Define our research design

- Key point: we know the exact probability distribution over Ω, and hence **D**.
- First consider with full knowledge for the true draw of **D** (the assignment that happened in our data)
- The fundamental problem of causal inference binds
  - Now, if we enforce that 50% is always treated, we know that there are only $\binom{10}{5} = 252$ potential combinations (each equally likely).

D_i	$Y_i(1)$	$Y_i(0)$	Y_i
1	11.9		11.9
1	10		10
1	9.7		9.7
1	9.5		9.5
1	11.4		11.4
0		7.6	7.6
0		7.1	7.1
0		7.7	7.7
0		8	8
0		7.8	7.8

Return to our estimand of interest, $\bar{\tau}$

- We now need an estimator for $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau_i$
- We already know under random assignment that $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ identifies $E(\tau_i)$
  - Take the empirical estimator of this expression: $\hat{\bar{\tau}}(\mathbf{D}, \mathbf{Y}) = \frac{\mathbf{D}'\mathbf{Y}}{\sum_i D_i} - \frac{(\mathbf{1}-\mathbf{D})'\mathbf{Y}}{\sum_i (1-D_i)}$
  - Note that this expectation operator is well-defined from the objects we already know – only D is random, and we know its marginal distribution over the sample
  - Can show that under certain assumptions (random assignment is equal across Ω) that this estimator is unbiased.
    - We can also now construct tests for this estimator that are more efficient than model based versions in small samples

- Is it an unbiased estimator in this case?
- If we assume that assignment is completely equal, then let $\pi_1(\mathbf{D}) = n_t(\mathbf{D})/n$ be the share treated, and $E(\pi_1^{-1} D_i) = 1$.
- Then,

$$E(\hat{\tau}(\mathbf{D}, \mathbf{Y})) = E\left(\frac{\mathbf{D}'\mathbf{Y}}{\sum_i D_i} - \frac{(\mathbf{1}-\mathbf{D})'\mathbf{Y}}{\sum_i (1-D_i)}\right) \tag{1}$$

$$= n^{-1} E\left(\sum_i \pi_1^{-1} Y_i D_i - \sum_i (1-\pi_1)^{-1} Y_i (1-D_i)\right) \tag{2}$$

$$= n^{-1} E\left(\sum_i \pi_1^{-1} Y_i(1) D_i - \sum_i (1-\pi_1)^{-1} Y_i(0)(1-D_i)\right) \tag{3}$$

$$= n^{-1} \sum_i Y_i(1) E\left(\pi_1^{-1} D_i\right) - n^{-1} \sum_i Y_i(0) E\left((1-\pi_1)^{-1}(1-D_i)\right) \tag{4}$$

$$= n^{-1} \sum_i Y_i(1) - Y_i(0) = n^{-1} \sum_i \tau_i \tag{5}$$

Variance of $\hat{\tau}$

- The variance of $\hat{\tau}$ (based on the sampling variation in the random design) is known thanks to Neyman (1923)

$$\sigma^2_{\hat{\tau}} = \frac{1}{n-1}\left(\frac{n_t\sigma_0^2}{n_c} + \frac{n_c\sigma_1^2}{n_t} + 2\sigma_{0,1}\right) \tag{6}$$

where n_t and n_c are the number of treated and control individuals ($n_t + n_c = n$) and $\sigma_0^2, \sigma_1^2, \sigma_{0,1}$ are the variance of the potential control, treatment, and the covariance between the two.

- Unfortunately, $\sigma_{0,1}$ comes from the joint distribution of $\mathbf{Y}_0, \mathbf{Y}_1$, and so isn't directly knowable. Instead, we bound for a conservative estimate:

$$\hat{\sigma}^2_{\hat{\tau}} = \frac{n}{n-1}\left(\frac{\sigma_0^2}{n_c} + \frac{\sigma_1^2}{n_t}\right) \tag{7}$$

The payoff – thinking about inference

- Now consider a test of our estimator. Consider the following *strong* null hypothesis: $\tau_i = 0$ for all i.
  - Note, this is much stronger than our traditional hypothesis testing based on the estimator
- Given our data, we can calculate the full distribution of potential observed statistics we would see, as we vary D.
  - How? By imputing our missing values using the null hypothesis, and calculating the estimator if we randomly permuted the treatment labels
  - Since we are asserting the known missing values, we can reconstruct the full distribution
- This approach is *very* valuable in other settings (especially when treatments are very complicated). More next week.
  - Key downside: doesn't test for *average* effects

D_i	$Y_i(1)$	$Y_i(0)$	Y_i
1	11.9		11.9
1	10		10
1	9.7		9.7
1	9.5		9.5
1	11.4		11.4
0		7.6	7.6
0		7.1	7.1
0		7.7	7.7
0		8	8
0		7.8	7.8

The payoff – thinking about inference

- Now consider a test of our estimator. Consider the following *strong* null hypothesis: $\tau_i = 0$ for all i.
  - Note, this is much stronger than our traditional hypothesis testing based on the estimator
- Given our data, we can calculate the full distribution of potential observed statistics we would see, as we vary D.
  - How? By imputing our missing values using the null hypothesis, and calculating the estimator if we randomly permuted the treatment labels
  - Since we are asserting the known missing values, we can reconstruct the full distribution
- This approach is *very* valuable in other settings (especially when treatments are very complicated). More next week.
  - Key downside: doesn't test for *average* effects

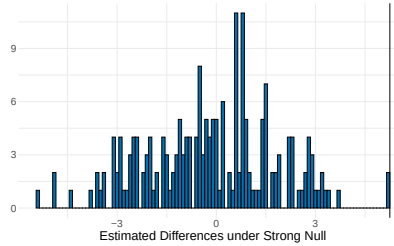
D_i	$Y_i(1)$	$Y_i(0)$	Y_i
1	11.9	11.9	11.9
1	10	10	10
1	9.7	9.7	9.7
1	9.5	9.5	9.5
1	11.4	11.4	11.4
0	7.6	7.6	7.6
0	7.1	7.1	7.1
0	7.7	7.7	7.7
0	8	8	8
0	7.8	7.8	7.8

The payoff – thinking about inference

- Now consider a test of our estimator. Consider the following *strong* null hypothesis: $\tau_i = 0$ for all i.
  - Note, this is much stronger than our traditional hypothesis testing based on the estimator

- Given our data, we can calculate the full distribution of potential observed statistics we would see, as we vary D.
  - How? By imputing our missing values using the null hypothesis, and calculating the estimator if we randomly permuted the treatment labels
  - Since we are asserting the known missing values, we can reconstruct the full distribution

- This approach is *very* valuable in other settings (especially when treatments are very complicated). More next week.
  - Key downside: doesn't test for *average* effects



Alternative estimator? Horvitz-Thompson

- For our estimator of $\bar{\tau}$, the estimator is unbiased only under certain assumptions (random assignment is equal across Ω).

- A more general approach is more flexible and unbiased in many designs, from Horvitz-Thompson (1952) (see Aronow and Middleton (2013) for a useful discussion):

$$\hat{\bar{\tau}}_{HT} = n^{-1}\left[\sum_i \frac{1}{\pi_{1i}} Y_i D_i - \frac{1}{\pi_{0i}} Y_i (1 - D_i)\right], \tag{8}$$

  where $\pi_{i1} = Pr(D_i = 1)$, and $\pi_{0i} = Pr(D_i = 0)$.

- This estimator is unbiased even in settings where we don't have equal weighting across the sampling space
  - This is reweighting using the propensity score!

Ok, great, but what's the problem?

- Inference in this setting is very agnostic to a broader sample
- How to think about extensions to other problems?
- More generally, does a focus on *internal validity* suffer from focusing too little on *external validity*
- This debate erupted at the end of the 2000s, especially focused on development
  - "Instruments, Randomization, and Learning about Development" Deaton (2010)
  - "Comparing IV with structural models: What simple IV can and cannot identify", Heckman and Urzua (2009)
  - "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" Imbens (2010)
  - "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy" Heckman (2010)
- Much of this is tied to instrumental variables, which we'll revisit later

“Instruments, Randomization, and Learning about Development” Deaton (2010)

In section 4 of this paper, I shall argue that, *under ideal circumstances*, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us “what works” in development, to design policy, or to advance scientific knowledge about development processes.

Project evaluations, whether using randomized controlled trials or nonexperimental methods, are unlikely to disclose the secrets of development nor, unless they are guided by theory that is itself open to revision, are they likely to be the basis for a cumulative research program that might lead to a better understanding of development. This argu-

“Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy” Heckman (2010)

TABLE 2
COMPARISON OF THE ASPECTS OF EVALUATING SOCIAL POLICIES THAT ARE COVERED BY THE NEYMAN–RUBIN APPROACH AND THE STRUCTURAL APPROACH

	Neyman–Rubin Framework	Structural Framework
Counterfactuals for objective outcomes (Y_0, Y_1)	Yes	Yes
Agent valuations of subjective outcomes (I_D)	No (choice-mechanism implicit)	Yes
Models for the causes of potential outcomes	No	Yes
Ex ante versus ex post counterfactuals	No	Yes
Treatment assignment rules that recognize the voluntary nature of participation	No	Yes
Social interactions, general equilibrium effects and contagion	No (assumed away)	Yes (modeled)
Internal validity (problem **P1**)	Yes	Yes
External validity (problem **P2**)	No	Yes
Forecasting effects of new policies (problem **P3**)	No	Yes
Distributional treatment effects	No[a]	Yes (for the general case)
Analyze relationship between outcomes and choice equations	No (implicit)	Yes (explicit)

[a] An exception is the special case of common ranks of individuals across counterfactual states: “rank invariance.” See the discussion in Abbring and Heckman (2007).

Ok, great, but what's the problem?

- Many of the complaints by the anti-randomistas devolve into three types:
  1. These are done incorrectly (e.g. bad IVs) – this is not interesting and bad research should be rejected regardless. More importantly, the transparency of the design should make this easier
  2. Inablility to generalize to other populations – e.g. Progressa is a big success, but knowing that conditional cash transfers work in this one setting does not necessarily inform our ability to roll it out in places that are very different
  3. A rhetorical overreliance on RCTs as the gold standard – post-hoc analyses (w/o pre-analysis plan) defeat the underlying value of an RCT anyway

- The concern is that this focus on RCTs and IVs causes an overfocus on irrelevant or unimportant questions. A briefcase full of results that are not economically useful.

My take

- My (biased) take on this:
  1. These concerns about empirics being too separated from models are overstated. Perhaps in part in response to these critiques, many empirical papers with causal parameters are tightly linked to theory models. For those that are not, they inform many theoretical papers. A push to open data has actually made it easier for researchers to follow-up and study these issues
  2. This concern about how to do empirical work does not provide much of a counterfactual (the counterfactual of the counterfactuals!). Evidence suggests that empirical work was in a not-so-great place historically.

- Most importantly: *the inclusion of an economic model does not grant an empirical researcher to omit a research design from their empirics*

- Many researchers may propose a model, and then demonstrate that their model is consistent with observational data:
  - This is a research design that needs to be made explicit