Metrics B Notes

by *Gabriel Chaves Bosch*

1 Overview of Topics and the Identification Problem

1.1 Angrist and Krueger (2001). IV and Search for Identification

IV were **originally used for identifying supply or demand curves**. Wright argued for the use of "curve shifters" that could allow us to identify changes in either supply or demand. Wright used different curve shifters for his demand/supply elasticities estimation, and then averaged the results, but the way in which information is combined in the most efficient way is, usually, two-stage least squares[1].

One of the other reasons to use IV is **measurement error**, for instance due to statistical agencies shortcomings or mismatches between variables in theoretical models and real-world measurable variables. An example of this?

But **IV has a problem: biasedness**. Consistency is not so much a problem but biasedness should be overcome with using large samples. Dig more on this.

IV and Omitted Variables

A flowering recent work uses IV to overcome OVB in estimating causal relationships. If we could held OV constant, we should not have to do this. But economic theory does not usually specify all the variables that should be held constant when estimating a given causal relationship, and it is difficult to measure all of the relevant variables even if they are specified.

One solution to OVB is to use **random assignment**. To see how IV can solve OVB, suppose we want to measure the cross-sectional rate of return to schooling ρ:

$$Y_i = \alpha + \rho S_i + \beta A_i + \varepsilon_i$$

Y_i is log wage, S_i is the highest grade of schooling completed and A_i is a measure of ability or motivation. A_i is usually unobservable. Without information on A_i, ρ is unidentified and thus cannot be deduced from the joint distribution of earnings and schooling alone.

Suppose we find Z_i which is correlated with schooling but unrelated to earnings. Therefore $Z_i \perp \varepsilon_i$. Then an instrumental variables estimate of the return to schooling is:

$$\hat{\rho} = \frac{Cov(Y_i, Z_i)}{Cov(Z_i, S_i}$$

IV allows us to estimate ρ consistently and free from OVB asymptotic bias, without knowledge or information of the omitted variables. One of the source of such IV's are natural experiments, which has generated some of the most provocative empirical findings in economics, along with some controversy over substance and methods.

A good instrument is correlated with the endogenous regressor for reasons the researcher can verify and explain, but uncorrelated with the outcome variable for reasons beyond its effect on the endogenous regressor. Good instruments often come from detailed knowledge of the economic mechanism and institutions determining the regressor of interest.

An **interesting example** of natural experiments is Angrist and Krueger (1991) quarter of birth IV. Men born earlier in the year tend to have lower average schooling levels. Because an individual's DOB is probably unrelated to the person's innate ability, motivation or family connections, DOB should provide a valid instrument for schooling. Older cohorts tend to have higher earnings, because earnings rise with work experience. In this paper, OLS and IV results are pretty similar though. This finding suggests that there is little bias from omitted

[1] Control variables should be present in both stages.

ability variables in the ordinary least squares estimate of the effect of education on earnings. Thus IV may be weak.

An **example in which OLS and IV differ substantially** is Angrist and Lavy (1999). This paper estimates the effects of class size using bureaucratic rules that create sharp discontinuities in average class sizes in Israel. OLS finds no effects while IV finds sizeable beneficial effect of smaller class size.

IV approaches contrast favourably with studies that provide detailed but abstract theoretical models, followed by identification based on implausible and unexamined choices about which variables to exclude from the model and assumptions about what statistical distribution certain variables follow.

Interpreting Estimates with Heterogeneous Responses

Sometimes, not every observation is affected by the instrument. Therefore, IV provide an estimate for a specific group–namely those people whose behaviour can be manipulated by the instrument. Hence, the effect in natural experiments is estimated for subjects who will take the treatment if assigned to the treatment group, but otherwise not take the treatment. This parameter is known as the Local Average Treatment Effect (LATE).

The author's view is that instrumental variables methods often solve the first-order problem of eliminating omitted variables bias for a well-defined population. Since the sample size and range of variability in many empirical studies are quite limited, extrapolation to other populations is naturally somewhat speculative and often relies heavily on theory and common sense. Moreover, the existence of heterogeneous treatment effects would be a reason for analyzing more natural experiments, not fewer, to understand the source and extent of heterogeneity in the effect of interest.

Potential Pitfalls

The most important potential problem is a bad instrument. IV estimates with weak instruments tend to be centred on the corresponding OLS estimates.

1.2 Duflo, Glennerster and Kremer (2007): Using Randomization in DevEcon Research: A Toolkit

Many of the randomized evaluations that have been conducted in recent years in developing countries have had fairly small budgets, making them affordable for development economists. Working with local partners on a smaller scale has also given more flexibility to researchers, who can often influence program design. As a result, randomized evaluation has become a powerful research tool.

The chapter thus provides practical guidance on how to conduct, analyze, and interpret randomized evaluations in developing countries and on how to use such evaluations to answer questions about economic behavior.

The Problem of Causal Inference

We have to make counterfactuals. To fix ideas, think of a *potential outcome*, introduced by Rubin (1974). Suppose Y_i^T is average test score of a kid i in a school with textbooks and Y_i^C of same kid i in a school without textbooks. If we are interested in the average effects of textbooks in the population, we are interested in

$$\mathbb{E}[Y_i^T - Y_i^C]$$

In a large sample, we can take differences between schools with textbooks and schools without:

$$D = \mathbb{E}[Y_i^T|\text{School has textbooks}] - \mathbb{E}[Y_i^C|\text{School has no textbooks}] = \mathbb{E}[Y_i^T|T] - \mathbb{E}[Y_i^C|C]$$

Subtracting and adding the non-observed quantity $\mathbb{E}[Y_i^C|T]$, we obtain:

$$D = \underbrace{\mathbb{E}[Y_i^T - Y_i^C|T]}_{\text{ATET}} + \underbrace{\mathbb{E}[Y_i^C|T] - \mathbb{E}[Y_i^C|C]}_{\text{Selection bias}}$$

The first term on the RHS is the effect of the treatment on the treated: on average, in treated schools, what difference did the books made. This is our quantity of interest. The other term is the selection bias, which captures systematic differences between schools with textbooks and schools without.

Since $\mathbb{E}[Y_i^C|T]$ is not observed, it is in general impossible to assess the magnitude (or even the sign) of the selection bias and, therefore, the extent to which selection bias explains the difference in outcomes between the treatment and the comparison groups. An essential objective of much empirical work is to identify situations where we can assume that the selection bias does not exist or and ways to correct for it.

One instance in which selection bias can be removed is when individuals are randomly assigned to treatment or comparison. Under randomisation, the ATE can be estimated as the difference between the two groups sample means,

$$\hat{D} = \bar{Y}^T - \bar{Y}^C$$

which converges to D asymptotically. If, in addition, the potential outcomes of an individual are unrelated to the treatment status of any other individual (Stable Unit Treatment Value Assumption, SUTVA), we have:

$$\mathbb{E}[Y_i|T] - \mathbb{E}[Y_i|C] = \mathbb{E}[Y_i^T - Y_i^C|T] = \mathbb{E}[Y_i^T - Y_i^C]$$

which is the causal parameter of interest for treatment T. The regression counterpart to obtain $\hat{D}$ is:

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

One can easily shown that $\hat{\beta} = \bar{Y}^T - \bar{Y}^C$

However, before proceeding further, it is good to note that the first expression,

$$\mathbb{E}[Y_i^T - Y_i^C]$$

estimates the overall impact of a particular program on an outcome. It may be different to the impact of textbook on test scores keeping everything else constant, and therefore there may be some general equilibrium effects (the total derivative discussion[2]) that imply that $\mathbb{E}[Y_i^T - Y_i^C]$ and $\mathbb{E}[Y_i^T - Y_i^C|X_i]$ may (substantially) **differ**, where X_i is a vector of demographic characteristics (we assume cross-section).

Other Methods to Control for Selection Bias

Sometimes, we can create comparison groups that are valid under a set of identifying assumptions. These are:

1. **Controlling for Selection Bias by Controlling for Observables**. The first possibility is that, conditional on a set of observable variables X, the treatment can be considered as good as randomly assigned. That is,

   $$\exists X : \mathbb{E}[Y_i^C|X, T] - \mathbb{E}[Y_i^C|X, C] = 0$$

   This is true when treatment is randomly assigned conditional on a set of variables X. This can be done through fully non-parametric matching or through propensity score estimates. However, the conditional independence assumption is usually not testable and most of the time researchers just include the other variables in the dataset, leaving much potential for OVB.

2. **RDD**. A very interesting special case of controlling for an observable variable occurs in circumstances where the probability of assignment to the treatment group is a discontinuous function of one or more observable variables. If the impact of any unobservable variable correlated with the variable used to assign treatment is smooth, the following assumption is reasonable for a small ϵ:

   $$\mathbb{E}[Y_i^C|T, X \in N_\epsilon(\bar{X})] = \mathbb{E}[Y_i^C|C, X \in N_\epsilon(\bar{X})]$$

   where X is the underlying variable and $\bar{X}$ is the threshold for assignment. This implies that for some ϵ-neighborhood of $\bar{X}$, selection bias is 0. The idea is to estimate the treatment effect using individuals just below the threshold as a control for those just above. This method is not used so much in DevEcon because in developing countries cutoffs are usually not totally enforced.

3. **DiD and FE**. If we have data before and after the treatment, stuff becomes easier. We just compare groups before and after (subtract the evolution of the control from the evolution of the treated over some fixed time period). The estimate is unbiased under the assumption that if the treatment had not been implemented, the outcomes would have followed parallel trends.

[2] If we let $Y = f(I)$, where I are different inputs to a test score, where we might be talking of a structural relationship.

External Validity and Generalizing Randomized Evaluations

While internal validity is necessary for external validity, it is not sufficient. Here we review threats to external validity of randomised evaluations and ways to ameliorate them.

1. **Partial and General Equilibrium**. Because randomized evaluations compare the difference between treatment and comparison populations in a given area, they are not able to pick up general equilibrium effects. Such effects may be particularly important for assessing the welfare implications of scaling up a program. General equilibrium effects of this kind can be thought of as another variety of externality.

2. **Hawthorne and John Henry Effects**. When the subjects change their behaviour due to the knowledgement that they are being treated. One way to correct for this is to use longer run data.

3. **Generalizing beyond Specific Programs and Samples**. Three factors can affect the generalizability of randomized evaluations: how the experiment is performed, the specific sample chosen and the reason to implement the program. Unfortunately, evidence on program replication is limited at this point. Ideally, institutions should emerge that both carry out such evaluations and ensure the discussion of the results to policymakers, even if academic publications are not the ideal forum.

1.3 Handout on Angrist and Krueger (2001)

1. **(Usage) What are the four key usages of instrumental variables?**

   - Measurement error. In the Y variable we have no bias, but in the X variable we have attenuation bias[3]. Regarding variance, measurement error increases the variance, creating noise in the estimates. If IV is actually correlated to the underlying variable and uncorrelated to the source of measurement error.
   - Selection bias.
   - Simultaneous equation.
   - Incomplete compliance. Related to intention to treat.

2. **(Definition and Theoretical Properties) How are IV methods constructed? What is the difference between small sample properties of OLS and IV? What does it imply for the usage of IV compared to OLS?**

   We usually have a independent variable x that might have some problems of endogeneity with respect to a dependent variable y, and we might want to "isolate" the exogenous variation of x using a instrumental variable z that is unrelated to the dependent variable y, so that is independent of the error term in usual regression settings. We can use either IV (simple) or 2SLS. For IV, we just use as coefficient

   $$\frac{Cov(Z,Y)}{Cov(Z,X}$$

   For 2SLS, we first regress X on Z, and obtain

   $$\hat{\delta} = (Z'Z)^{-1}Z'X$$

   With this, we obtain

   $$\hat{X} = \underbrace{Z(Z'Z)^{-1}Z'}_{=P_z'}X$$

   and then just run a regression of Y on $\hat{X}$. Therefore,

   $$\hat{\beta}^{2SLS} = (X'P_zX)^{-1}P_zY$$

[3] With measurement error in the RHS, if we create variation in the RHS without creation variation in the LHS, then we might be effect of X on Y towards 0

where we used symmetry and idempotency of P_z.

Regarding small sample properties of OLS and IV, we have that IV is biased in small samples, specifically if the instrument is weak. Since IV is consistent, in order to overcome biasedness, we should aspire to work with large samples.

3. **Questions on Angrist and Krueger (1991)**

   (a) **What is the IV they used and what is their story and the institutional settings that motivated their choice of IV?**

   Dropout age is based on the birthday, but entry depends on the calendar year in which they turn 6. Kids born in December 31st are eligible to school with a younger physical age.

   Recall that IV has to have relevance and exclusion. These two are shown in the paper to hold, while exclusion is usually very difficult to show.

   (b) **In their data, the IV estimate of the return to schooling is not very different from the OLS estimate, while the two estimators often yield very different estimates in other settings? What could potential reasons?**

   The authors argue that selection bias might not be that high. The estimate is a LATE, using only people that really wanted to drop out, so that's a small proportion of the population. Also, if we allow heterogeneity in β it may be the case that β for this particular group might be the overall β for this population.

   However, a lot of people do not trust IV. When we use IV we have to think how much the sample we have is representative of the overall population.

   (c) **What is the specific group of people affected by the treatment in their analysis? How does it affect the interpretation of results?**

   The subpopulation affected to IV is very small. But these dropouts are policy-relevant population.

4. **Redo the previous for Angrist and Levy (1990)**

   Wage earning using military lottery as IV.

5. **(Issues in the Interpretation) What are common criticisms of IV methods? What are the justifications?**

   One is that is external validity. Another is that we do not understand the channels between education and earnings. Structuralists would tell us that we need to know what is the choice process.

6. **(Pitfalls of IVs) What are the typical pitfalls in using IVs?**

   Weak IV, where correlation is weak. Bad IV, where exclusion restriction is likely to be violated. This last one depends pretty much on the argument.

1.4 Handout on Duflo et al (2007)

1. **Do estimates of treatment effects in randomized evaluation capture partial effects or total effects? How does it affect welfare evaluation based on randomized evaluation?**

   Suppose the outcome is a function of given inputs:

$$Y = f(I) = f(I_0, I_1, \ldots, I_K)$$

   The partial effect for a given input k is

$$\frac{\partial Y}{\partial I_k}$$

   The total effect would be:

$$\frac{dY}{dI_k}$$

   Given the reoptimization, we will be estimating the total effect, because changing one of the inputs makes the other inputs be changed given the reoptimization of the agents.

Regarding welfare, since the RCT causes a overall change in the different variables in the reoptimization, we usually cannot evaluate the welfare change due to the RCT. This is one of the things that structuralist like: in structural estimation, welfare can be measured as utility minus costs, while in reduced form econometrics this cannot be precisely done.

2. **What are the other typical ways of dealing with selection bias? What are their drawbacks?**

   It is the syllabus. Except matching. But matching is not that credible. Matching usually gives bad estimates. However, it is better than just OLS, because we are making a non-parametric comparison of the group, instead of assuming linearity and failing prey of misspecification of the functional form. Matching uses a less restrictive functional form and we have less attenuation bias. But still it has been proven that matching is not that reliable compared to RDD and related methods.

   Normally, we need a identifying assumption to tear apart causation from correlation. In RCT's, this is

   $$E(Y^C|T) - E(Y^c|C) = 0$$

   In OLS, we estimate:

   $$E(Y^C|T, X) - E(Y^c|C, X) = 0$$

   This approach is mostly problematic. Most of the time it is almost impossible to control for unobservables (unmeasurable variables)[4]. Also, another important assumption of OLS is linearity, and we might not be approaching the correct functional form. Finally, we might have theoretical incompleteness of what goes in the RHS.

   We can do RDD (explained above).

3. **Do randomized evaluations capture general equilibrium effects of a program? How does this issue affect the internal and external validity of the results?**

   RCT's capture partial equilibrium effects. However, general equilibrium effects might cause spillovers that might be a threat to internal validity of the findings. An example of a general equilibrium effect would be when in the implementation of a program, the treated affect the control. When the experiment is small, the general equilibrium effect might not be present and internal validity might not be threatened.

   In a nutshell, if the experiment is pretty large, the general equilibrium effect is going to be pretty present, and then the threat to external and internal validity will exists. When a small experiment is performed only the partial equilibrium effect might be captured, so internal validity is satisfied but unfortunately if the experiment was to go bigger then the external validity is highly threatened.

4. **Discuss examples in which Hawthorne effects and John Henry Effects can be present. How can we address those problems?**

   Hawthorne effects arise when the treated are affected by the treatment itself. John Henry Effect is the same but for the control group. For instance, when experiments are artificial, the individuals change the behaviour due to the experiment.

   We can either have long-term observations or not tell them about the experiment.

5. **What are the typical factors that affect the generalizability of randomized evaluation results?**

   Answer

2 Class 1 Notes

2.1 Reduced form vs structural estimation

Reduced form is distinction between causation and correlation. We want to recover underlying causal effects in the data. In structural, we want to test the causation that we obtain from theoretical models.

[4]Check also correct way of clustering paper by Duflo

In reduced form, we have the typical model

$$y = X\beta + u$$

Most of reduced form econometrics is about distinguishing/identification the causation from the correlation of X on y. β here is the reduced form parameter.

In a structural model we have structural parameters that are defined by choice process. In a full blown structural model we have a model of a optimisation process. We do not have the actual outcome of the model yet in structural information. In a sense, structural estimation is closer to ex-ante evaluation while reduced form is closer to ex-post evaluation of a problem, with the given differences of the problem itself across techniques.

Structural estimation and structural calibration of parameters is not the same, but structural estimation is closer to structural calibration than it is to reduced form econometrics.

In the structural model we usually want to write down the primitives (ingredients of choice process that do not change whatever is the environment):

- Preference of consumers.
- Cost function of producers.

Keane and Wolpin: what actually happens if they change the cost of the college? using a structural model.
The four uses of IV are:

- Selection Bias
- Simultaneous equations
- RCT (Compliance)
- Measurement error in the RHS.

2.2 IVs

In IV, if we have the same dimensionality, we have:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'Y$$

If we have multiple instrumental variables, we can use the 2SLS estimator. The best possible way to aggregate many instrumental variables is to run first a regression of X on Z, and then with the estimate $\hat{X}$, we would have the following estimator

$$\hat{\beta}^{2sls} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

However, there are problems with IV. There are two problems with weak instruments: strongly biased in small samples and very big SE (not relevant enough). But the worst is bad IV, which are also maybe endogenous. Sometimes we have to deal with how persuasive the history behind the IV is.

2.3 RCT

RCTs were motivated by the potential outcome framework. This is explained in Duflo's summary of the paper above.

Recall one of the facts of RCTs is that **total effects** are captured. That is, if our outcome is formed by a set of inputs,

$$Y = f(I_0, \ldots, I_K)$$

Then we capture dY/dI_0 instead of $\partial Y/\partial I_0$.

3 Experiments

3.1 Moffitt (2004): The Role of Randomized Field Trials in Social Science Research

The article deals with the assessment of the usefulness of RFTs (randomised field trials) in welfare analysis, and outlines its methodological and institutional drawbacks. Up to the writing of the paper in 2004, the area of social welfare is one that has seen the most intense use of RFTs. Moffitt focuses on reforms to the Aid to Families with Dependent Children (AFDC) programme in the US for the 1960-2000 period, and its related evaluations.

AFDP provides cash benefits to low-income families where children are present and one biological parent is absent from the household. In the 1950-60's the programme lost popularity but families with single mothers started to increase, so representatives became interested in reforming the programme in a variety of ways, focusing mostly on increasing the work levels of the recipients. This pressures meant than in the 1960-70's, reforms generally took the form of providing financial incentives by lowering the benefit-reduction rate, or tax rate, in the AFDC programme. Therefore, in the 1970's, policies shifted to work requirements instead of financial incentives to make people work more, with also a focus on training and education. But in the late 80's work levels remained low amongst AFDC participants and thus in the 1990's requirements for work became more stringent. In the late 90's, the program was replaced by the Temporary Assistance for Needy Families (TANF) programme.

Regarding field experiments, the ones that stand out the most are the Negative Income Tax (NIT) experiments. The control group in these experiments received the AFDC, for single mothers, or no programme for the rest of household. The objective was to assess whether reduced tax rates increased work levels, but it was surprisingly not the case.

Critics of the NIT outlied problems in the randomisation across control and treatment groups, as well as unequal attrition rates across groups. Analysis methods were also critised. In general, these shortcomings come from the side of researchers, by poor design of (some of) the experiments. These 70's NIT experiments were proceeded by many RFTs in the 80's.

In the 80's the federal government allowed states to experiment using tweaks in the AFDC programme. Indeed, in the late 80's, experimentation was on the verge of being mandatory. Therefore, the experiments of the 1980s did not have the design, attrition, or analysis weaknesses of the NIT experiments. These experiments were very simple and main outcome variables related to labour supply elasticities. As a result of these strenghts, the 1980's experiments were highly influential.

Welfare policy took a sharply different direction in the 1990s with the introduction of stronger work requirements, sanctions, and time limits. The RFTs followed this shift. The scope of these RFTs lies somewhere in between the modest AFDC experiments of the 1980s and the larger scale experiments of the 1970s. Although they have tested reforms that are more far reaching than the incremental reforms tested in the 1980s, they fall far short of the radical reform tested by the NIT. In most other respects, however, they are similar to the RFTs of the 1980s, and they share similar levels of credibility.

There have been relatively few new RFTs begun since 1996. The primary reason is that the 1996 law devolved the program to the states and, hence, most federal regulatory authority disappeared as well because the states are no longer required to design programs according to any particular structure dictated by federal law. Experiments ain't needed anymore.

3.1.1 Strenghts and weaknesses

Strengths are that RFTs have been professionally conducted and analysed, and that randomisation was properly carried. Internal validity thus is pretty strong. The advantages of simplicity of design, use of administrative data, simplicity of analysis methods, and policy relevance have led to a set of quite credible policy impact estimates.

However, the author discusses some of the weaknesses specially pertinent to cash welfare RFTs. These are the following:

- **Contamination of control groups when estimating the effects of systemwide reform**. RFTs do not pick up the feedback or macroeffects that would occur if an experiment programme were implemented nationwide. One instance are general equilibrium effects that propagate through markets. For instance, a large supply of individuals to a particular labour market due to the nationwide implementation could change equilibrium prices or unemployment rates, which would then feed back and alter the behaviour

of individuals in the population, generating an effect that is not captured by small-scale RFTs. This is however not a threat on external validity. However, the cash welfare RFTs of the 1990s were vulnerable to a more serious manifestation of this problem because they took place in an environment in which such macroeffects were actually occurring and that almost certainly affected the outcomes of control group members. This is a more serious problem because it affects internal validity rather than external validity. One key reason is that the nature of experimentation gradually shifted over the 1990s from small-scale to large scale RFTs.

- **Inability to estimate entry effects**. Entry effects occur when the implementation of a programmatic reform in an existing program alters the rate at which individuals apply for that program or gain admittance to it through the selection process of program operators. There can be either direct or indirect effects. Direct effects can occur when the reform actually involves a change in the "front door" admission process by which applicants are handled. Indirect effects can occur when the policy reform affects the attractiveness of the program, in either a positive or negative direction, and consequently affects the rate at which eligible apply.

  RFTs to capture entry effects could be designed if the unit of observation were communities, or local areas, where a programmatic reform is offered in some areas and not others, because then the impact of the reform on the entry rate could be estimated by a comparison of that rate across experimental and control areas. Nevertheless, the problem of inability to estimate entry effects and participation rates in small-scale RFT tests is a problem only of external validity because it implies that the generalization and extrapolation of the experimental results to a national program would provide an incomplete estimate of its total impact, much in the same way that macro-, feedback effects are missed.

  A useful contrast to the recipient-only welfare RFTs of the 1980s and 1990s is provided by the NIT experiments, because those experiments differed from the later AFDC experiments by enrolling in the experimental and control groups a random sample of the entire low-income population in the area. Thus, individuals were enrolled who were not on welfare and who were, when randomized into the experimental group, offered the opportunity to enter the new welfare program if they wished but were not required to. Thus, entry effects—or, really, participation rate effects because this was a new program—were partially captured. Entry rate effects induced by changes in the welfare tax rate also were captured because the samples in each separate experimental cell included a random sample of the entire population, including nonparticipants, and thus, the effect of changes in the welfare tax rate on participation in the program could be estimated. This difference may be part of the explanation for why the estimated effects of welfare tax rate reductions in the NIT experiments on work levels and earnings showed no effects, whereas those of the 1980s and 1990s AFDC RFTs generally showed a positive effect.

- **Issues related to site effects**. The RFTs listed in Tables 1 and 3, including the NIT experiments, were conducted in a single area or a limited number of areas. The problem of external validity that this raises—that such areas may not be nationally representative and hence their results may not be a correct estimate of nationwide implementation—is a familiar one that has been discussed thoroughly in the literature. The problem arises if area-level characteristics interact with the impact of the treatment on outcomes and not so much if individual-level characteristics so interact.

  The RFTs of the 1980s and 1990s are superior in this respect to those of the 1970s because a greater number have been conducted and in a much larger number of areas. Unfortunately, however, the key problem in learning the interactive effects of area characteristics on treatment impact is that the area variation embodied in the RFTs was not planned in any systematic way to provide variation from which something could be learned.

- **Limited and unplanned treatment variation**.A rather related issue that, again, pertains to the ability to use RFTs to learn lessons for the future is the extent to which the RFTs of the 1980s and 1990s reflected limited and unplanned treatment variation. The analogous problem to area variation discussed previously was present in this case as well, because the programs tested in different areas did not vary particular treatment features while holding others fixed, thus preventing learning the incremental effects of particular treatment components. Once again, the political constraints on the constellation of RFTs that were conducted is one of the primary reasons for this lack of variation.

  The experience of the RFTs of the 1980s and 1990s in this respect poses a political difficulty for experimental design if the estimation of the incremental effects of individual components are of interest. The policy

makers over this period were initially not interested in testing the effects of individual treatment components added on top of the then-existing AFDC program. This is because the policy makers believed that the effects of the individual components interact and that the sum total effect of the bundle as a whole would be greater than the effects of any individual component-introduced piecemeal. Once again, it was the effect of transforming the program in a major way that was the object of interest.

- **Problems of black-box treatment designs**. The final issue is that of black-box treatment designs. Black-box treatments are those constituted of multiple complex treatment components that are either difficult to describe or that allow considerable discretion when implemented in the field. A welfare-to-work program, for example, which consists of some type of initial assessment of job skills, assignment to a type of work or training program for which the caseworker is given general guidelines but allowed discretion, followed by a sequence of work programs and sanctions, the latter of which is also partly at the discretion of the caseworker, is a case in point.

  Some analysts also regard black-box treatments as those where the mechanism by which the treatment has an effect is not understood or where there is no theory to guide the experiment. The problems that black-box experiments raise are, first, that they are difficult to replicate and, by extension, difficult to generalize to a national program; second, it is difficult to compare different black-box experiments to each other or to extrapolate from them to programs that may differ from them in small or large ways.

Each of the five limitations has a lesson. The first one is that RFTs are best used when they attempt to estimate the effect of incremental reform within a given, overall programmatic structure and are poorly designed to estimate the effect of system-wide structural reforms. Second, RFTs should be supplemented by nonexperimental analyses of entry effects where it appears possible that those effects are significant. Consequently, RFTs should be reserved for estimating the exit effects and effects on initial participant populations. The other three limitations discussed in this article all pertain to external validity and are concerned with ways to learn more about policy alternatives than recent RFTs have been able to do.

3.2 Gerber et al (2009): The Effect of Newspapers on Voting and Political Opinions

The experiment assigns random subscriptions to the Washington Post (liberal) or Times (conservative) amongst treated and nothing on controls. The main findings are no effect on voting turnover, knowledge or opinions. However, both newspapers switch to a higher support of democrats (More information makes you vote republican?). The experiment took place in 2005 before the Virginia governor race. The treatment group was administered either paper for ten weeks prior to the election, and a follow-up survey was made.

According to the authors, this paper is the first field experiment measuring the effect of newspapers on political attitudes and behaviour. The most important limitation of the study, though, is the small sample size.

The experiment studies individuals who do not already subscribe to a newspaper, hence are examining the effect of exposing individuals who, on average, are less exposed to the media than the average individual. Regarding the treatment, there were three noncompliance issues to note regarding treatment administration. 6% of the households opted out of the free subscription. In the analysis the authors focus on intent to treat effects and included all treatment group subjects even if they cancelled. Also, some addresses were undeliverable and some others already had a subscription (maybe only Sunday version).

Regarding the **outcome data**, the authors interviewed 1081 out of the 3347 in the sample for follow up. They asked questions about the 2005 Virginia gubernational election, national politics. Table A1 shows covariate balance and shows that while individuals who voted in 2002 and subscribed to a news magazine (hence are more engaged in politics), as well as those who preferred the Democratic candidate for governor in the baseline, were more likely to complete the follow-up phone survey, sample selection bias is not correlated with assignment to treatment.

One limitation of this study is that while we know which households received newspapers, we cannot be sure that the newspapers were read. Our follow-up survey provides three measures of the effect of newspaper provision on newspaper reading: whether subjects receive a newspaper, which newspaper they receive, and the frequency with which they read a newspaper. Treatment assignment and readership outcomes are significant in pooled regressions amongst newspapers, but the individual coefficients suggests less than full readership. Also, the authors say that some people actually subscribed to the Post after the treatment, pointing towards the fact that the newspapers were not disregarded.

Regarding the **results**, the newspaper did not make the people have more knowledge about politics. There was no effect on either self-reported or administratively measured turnout for the 2005 election. However, the newspapers had an effect on which candidate the subject supports. Getting the Post is estimated to increase the probability of selecting the Democrat by 11.2 percentage points. Contrary to initial expectations, the right leaning Times was also associated with an increase in the probability of a Democratic vote in the Virginia governor's race. And the differences between them are not statistically significant.

One of the explanations is the particular news environment, which was politically challenging for Republicans. It may be that what the coverage had in common was more important than any differences between the newspapers. Another explanation is that the Democrat candidate in Virginia back then was a conservative-leaning one. A third explanation could be sampling error: it is possible that there are meaningful differential effects, consistent with the news slant of the papers, which the authors did not detect due to inadequate power.

In general, the paper has a nice field experiment approach that can be applied in subsequent research.

3.3 Handout on Moffitt (2004)

First of all, note that the nature of experiments in this social welfare programmes, the treatment and control groups are usually much larger than in development RCTs.

Second, also note that in the 60's the negative income tax was designed so people would work: if you get 100 pounds a week working but 100 pounds not working, you will not work. You need to incentivise people to work for the welfare program to work. So, a rationale for NIT is making people actually work. Therefore, in the subsequent decades, a lot of training and education requirements had to met for eligibility.

1. **What are the five key limitations of experiments discussed in this paper?**

   (a) Contamination of control groups when estimating the effects of systemwide reform. It is also known as feedback mechanism.

   (b) Entry effects.

   (c) Site effects

   (d) Unplanned treatment variation.

   (e) Black-box treatment design.

2. **Discuss the mechanisms through which feedback effects affected internal and external validity of randomized field trials in the context of cash welfare programs.**

   The size of the entire experiment is what determines the problems. If the experiment is done in a very large scale, it's going to change everybody involved in that environment. Therefore, the control will be affected in a certain way and thus hinder internal validity.

   In case the experiment is done in a very small scale, the control group is likely to be exempt from any effect but then external validity might be threatened.

   For NIT experiments, some mechanisms could be for instance that the control group might reduce labour force participation due to higher competitiveness by the treated, who might demand more work.

   Sometimes, the feedback mechanism are encouraged from a social and political point of view.

3. **Discuss how the effect of a program depends on the entry effect. How did ignoring entry effects affect the evaluation of increased earnings disregards (reduced welfare tax rates on work levels)?**

   In experimental setting, we want to have a well defined and strict definition of the control and treatment groups. In the previous exercise, the treatment was not affected. However, entry effects can happen when people from control group want to be in the treatment group. Very often, large scale reforms contaminates the eligibility criteria. Some portion of the control group moves to treatment.

   In terms of outcome, the guys in the control group that go to treatment, we can ignore them, but then we are missing people. This is based on the assumption that the government can observe the entire situation.

   Depending on the unit of observation, we can solve this by doing randomisation based on communities or larger levels of aggregation. We can measure how the rate of participation changes across counties, and then we can see how the entry rate changes. Randomisation has to be balanced across counties, of course.

4. **How did site effects influence the program evaluation in cash welfare in 1980s and 1990s?**

   In a lot of situations, states have handfuls of counties, like Delaware or NJ, and therefore we have a small number of sites from which to compare. The intrinsic characteristics of the countries can be combined with the actual treatment effect.

   If you want to know how severe are site effects, you can run a regression of treatment outcomes on site traits. Then, we wont be able to overcome higher dimensionality. A fix to this, is to link observables of different communities to treatment status and see how site characteristic matters. But this is easier said than done due to the curse of dimensionality.

5. **Why was the problem of limited and unplanned treatment variation not prevented in cash welfare RFTs in 1980s and 1990s?**

   Governments usually impose many restrictions at the same time. Therefore it is difficult to disentangle the effect of continuous reforms and changes to the programmes. Or sometimes the reforms can be too heavy.

6. **What are common solutions to black-box treatment designs?**

   We want to reduce dimensionality of variation in treatment. And then measure them in a certain way, for instance caseworkers measured by skill. A way is to introduce fixed effects of caseworkers.

3.4 Handout on Gerber et al (2009)

Media slant is defined as the implicit bias of a certain media company.

Lim: why do OLS is not yielding good results? Because there is a correlation between using media and their voting. This is the main motivation for this paper.

1. (**Site Effect**) **This study is based on 2005 Virginia gubernatorial election. Discuss how the nature of this election may have affected the results of this study. That is, what are the specific aspects of the results that seem to be driven by the way that this race went?**

   The place is kinda close to DC. It might be a special place in terms of politics. That many people do not participate in politics is not enough: more characteristics should have been assessed, in demographic terms.

   In US politics there is a coat-tail effect. During this experiment, president Bush's Iraq war was affecting the tendence towards voting democrat.

   Lim advisees us to google about this election. This election involves Tim Kaine, who become then very popular. For instance, Tim Kaine was against death penalty.

   Media slant did not matter, but media exposure mattered. This might be driven by the Tim Kaine effect.

2. (**Sampling**) **Discuss how the sampling choice in this study may have affected the results.**

   There was a problem of compliance, and thus a small effect, as seen in the little knowledge of current events (lol).

   Also, the number of treated was small, and standard errors are quite high. Some results are driven by the size of the SE rather than the smallness of the point estimates in some cases. This ended up driving null results while probably there was some effect which was not precisely estimated.

   Moreover, the sample was about people who live close to DC but that seem to be more disengaged with politics, at least national ones. Given the sample, it was likely that the experiment would have had a very small treatment effect.

   Furthermore, if there is compliance issues, authors could have run an IV with intention to treat.

3. (**Interpretation of the results**) **The authors report that both The Washington Post and The Washington Times increased the support for Democrats and argue that media exposure was more important than media slant. Is this interpretation fully convincing? If not, elaborate the reasons.**

4. (**Interpretation of the results**) **The authors report clear compliance issues. Are there ways you would analyze and interpret results differently to address compliance issues better? Elaborate your ideas.**

   The treatment did not happen in some cases. Some people got the paper and did not read it. Some people opted out.

   Given that intention to treat: Some people were coded as treated but did not comply with the actual treatment. Some people are mistaken as treated, but are actually controls. The actual treated group is smaller than what we have. Therefore, we have a downward bias.

   They could have solved as IV. Suppose $X =$ treatment status and Z is ITT, where the former is the actual status and ITT is the status by the researcher. In this study they could have done this because they know X. But for some reason, they did not.

5. (**Broad implication; Experiment vs. observational study**) **The authors argue that conducting the first experimental study on the media influence on political attitudes, knowledge, and voting behavior is a contribution. Are there any limitations on the broad implication of this study caused by focusing on providing free subscriptions to newspapers? How does the nature of media influence in this study differ from that in other studies (any studies you know)?**

   Subscription is maybe not the best way to expose people to media. Free subscriptions maybe was something that people did not want. It is a weak mode of treatment.

3.5 Review

- Experiments capture the total effect instead of the partial effect.
- However, only partial equilibrium effects are captured instead of general equilibrium effects.
- Therefore GE effect and feedback effects harm external validity, due to effects mostly on the control group.
- Also, internal validity also depends much in the size of the experiment. If we have a large scale experiment, like in the welfare problem experiment, we might have violate internal validity.
- In the case of a small scale experiment, as the ones by NGOs, internal validity is typically valid but external validity is likely to be threatened.
- Moffitt addressed some of the threats to the experimental settings of the welfare problems. The first one was the feedback effects. The second was contamination due to entry effect (which can be solved by randomisation at community level instead of individual level, and then we can measure the entry effect). The third was the site effect, which is directly related to external validity. We have for instance endogenous spatial sorting (people go there during the experiment for some reason). Another could be for instance some effects given the place where experiment takes place, not necessarily because population for endogenous spatial sorting. The fourth was limited (or unplanned) treatment variation, which arises when we have too much changes in the RHS because the specification of the reform is multicollinear. The fifth was the blackbox treatment, in which the treatment is too weird and complex to identify the causal effect, or to replicate the experiment.

4 Natural Experiments

4.1 Duflo and Chatto (2004). Women as Policy Makers

This paper aims to study women's leadership on policy decisions. To do this, the authors explore political reservations for women (where women are required to fill certain government jobs). The main results are that reservation of a council seat affects the types of public goods provided, given the needs of the gender of the policymaker.

Previous research did not shed light on the causal effect of women's representation on policy decisions. Even if we know more about this effect, we would not know about the effect of quotas to enforce greater participation of women in politics. Therefore, Duflo's paper studies the policy consequences of mandated representation of women.

The identification strategy relies on the randomisation of the Indian government on village representation by women, specifically from Udaipur in Rajasthan and Birbhum in West Bengal, and compare investments made in reserved (for women) GPs (Gram Pradhan, which are village-level representatives) and unreserved GPs. Reservations were random. The gender preferences of men and women are proxied by the type of formal requests brought to the GP by each gender. The experiment confirms differences in GPs reserved for women, specially water sanitation and roads investments (for this last, comparatively more in Birbhum but less in Udaipur). Robustness checks confirm that the impact of reservation is driven only by gender of the Pradhan.

This paper is consistent with the following paper, Pande (2003), which shows that in Indian States where a larger share of seats is reserved for minorities in the State Legislative Assembly, the level of transfers targeted towards these minorities is also higher.

GPs are village authorities that encompass around 10.000 people in several villages. GPs do not have juridisction over urban areas, so this is all rural stuff. The major responsibilities of the GP are to administer local infrastructure and identify targeted welfare recipients. The GPs have also no control over teachers or health workers, but in some states there are Pnachayat-run informal schools.

The two states choosen differ quite a lot in terms of the GPs. West Bengal imposed them quite a long ago, in the 70's due to the communist party present back at that time, while Rajasthan implemented fully in 2000. Unreserved GP's have 6.5% women in West Bengal while only 1.7% in Rajasthan (only one woman).

The empirical strategy and data is the following:

1. **Data collection**: Two surveys were conducted. The first, in 2000, was conducted in Birbhum, a mainly agricultural region. As expected, given the random selection of GPs, there are no significant differences between reserved and unreserved GPs, with jointly insignificant differences.

   Regarding the data, the GP Pradhan was interviewed for demographic questions, and then two random villages plus the GP Pradhan residing village. Resources maps were also drawn with the help of villagers, asking them about when infrastructure was built. This method yields extremely accurate information about the village according to authors. Also, information about GP meetings was collected.

   Finally, in 2002, the same information was collected in Udaipur. The latter is much poorer than Birbhum, with lower literacy rates, but bigger villages with schools, health facilities and better road connection (this is why people there are not as concerned about roads maybe?).

2. **Empirical Strategy**. Thanks to the randomisation build into the policy, the reduced form effect of the reservation status can be obtained by comparing the means of the outcomes of interest in reserved and unreserved GPs. **But what we are trying to estimate is the effect of being reserved for a woman, rather than not reserved, in a system where there is reservation.**

   The main regressions have as outcome variables the outcome of interest for a given good, like investment in drinkin water between 1998 and 2000. The investment measured is standardised for the different categories of goods in both samples.

   Women elected as Pradhans differ from men in many dimensions. Controlling for their characteristics can be misleading because these can be endogenous to the reservation system.

The results are the following:

1. **Effects of political participation of women**: the fact that the Pradhan is a women significantly increases the involvement of women in the affairs of the GP in West Bengal. This does not happen in Rajasthan, but this is because women already participate more in the Gram Samsad in Rajasthan.

2. **Requests of Men and Women**. They differ.

3. **Effects of the Policy on Public Goods Provision**: Both in West Bengal and in Rajasthan, the gender of the Pradhan affects the provision of public goods. In both places, there are significantly more investments in drinking water in GPs reserved for women.

   These and other results in the paper suggest that the reservation policy has important effects on policy decisions at the local level. These effects are consistent with the policy priorities expressed by women.

   Finally, public good provision OLS regressions point toward the fact that individual women are not particularly more responsive to the needs of people in their communities. Rather, it is because their own preferences are more aligned to the preferences of women that they end up serving them better.

4.2 Mandated Political Representation for Disadvantaged Minorities

This paper exploits the institutional features of political reservation in Indian states to examine the role of mandated political representation in providing disadvantages groups influence over policy-making. The author finds that political reservation increases transfers to groups which benefit from the mandate. Parting from the previous paper mandated representation style of experiment, here it is similar but for minorities. Indian experiment is quite radical, as it makes it compulsory for some candidates to be minorities. This way, the effect of this policy (or mandate) is to alter legislator identity without affecting voting identity.

In other words, the Indian constitution requires that the extent of state-level political reservation enjoyed by a group reflect the group's population share in the state. This comes with a lag due to the low frequency o population census. This feature is exploited by the author to isolate the effect of reservation on policy outcomes. The main finding, that of increased redistribution toward the political reservation groups is accompanied by increases in overall spending and decreases in spending on education programs. Thus, reservation influences policy making and legislators belonging to minorities have used this influence to increase the incidence of targeted redistribution.

One of the main contributions is to shed light on how a country's choice of political institutions mediates the relationship between legislator identity and policy outcomes, specifically using as a particular institution political reservation.

4.2.1 Institutional Background and Data

Quantitative evidence on how such representation has affected electoral and policy outcomes is, however, lacking, and political commentators remain divided on this issue.

The empirical analysis of the paper exploits the diachronic variation in the extent of political reservation enjoyed by a group in a state. The cause of such variation is defined by Section 3 of Article 332 of the Indian constitution. Section 3 states that the proportion of jurisdictions reserved for scheduled caste (scheduled tribe) in the state. Moreover, the only permissible basis for changes is the extent of reservation enjoyed by a group in a state is changes in the census estimate of the group's population share in that state.

Data The unit of observation is Indian state. THe author uses data for the 16 major indian states, spanning 1950-1992. These states account for over 95 percent of the Indian population.

boring

4.3 Handout on Duflo's (2004) Women Politics

1. **(Importance of the Research Question) Why is their research question important? Discuss specific arguments the authors make to promote the importance of their research question.**

    - The paper is important because 30 countries and 12 EU countries had some kind of quota, but still women are left behind in politics. So it is a **specific and important policy question**. Generality of the people and the scope of the implication makes it actually relevant. Women are politically under represented: 14% female representative for 50% of the population: the problem is severe.
    - The paper is also important because it manages to disentangle a causal effect that was not possible in previous research. This is because in the given natural experiment by the Indian government the quotas are assigned randomly to localities and thus a counterfactual can be assessed.
    - The paper gives a theory with ambiguous conclusions, and the authors can answer the broader theoretical question with the causal effect. They motivate it because traditional canonical model predict that gender (or race) is unimportant, but this paper (or the next paper) show it is actually important theoretically.

2. **(Broad Implications) How helpful are the results from this paper in understanding other policies that foster women's political representation? Discuss specific aspects of the reservation policy studied in this paper that strengthen or limit its broad implications.**

    - The causal effect analysed in the paper is the causal effect of reserved compared to nonreserved units, but not the causal effect of having women.

- The Indian case has many peculiarities. For instance, in unreserved communities were like 1% women. But the main thing is the **cultural setting, specially gender rules**. What about the outcome variable, which is the name of complaints about roads and drinking water, which is only relevant in the economic development setting. Implementing these quotas are much more difficult to implement in legislative sits in Europe. Therefore **external validity is quite limited**.
- **One of the strenghts** of the paper is that they are super precise about the mechanism on how these mandate representation (women) affect policy outcomes. It can be the Pradhan or it can be that the Pradhan is aggregating preferences. This is related to question number 7.

The results point towards the fact that the causal effect of mandated representation alters the participation of women in politics, involving them more if there are more women politicians, and also noting that women policy makers tend towards policies aligned to gender-specific preferences. Therefore, to ensure a correct representation of the population given its demography, quotas on their characteristics help them be proportionately represented in practical terms.

The results enjoy internal validity due to the randomisation of the experiment and the correct specification and identification strategies. However, there are threats to external validity, particularly given the special composition of the GPs, mostly rural and relatively uneducated.

3. (**Alternative Research Design**) **Suppose that you design a research project of your own on the influence of women as policy makers, without natural experiments. What would be an alternative research design and data sets you would consider using?**

   An RCT with quotas proportional to the demographic composition of genders, where the deviations from the pre-intervention composition of genders in policy would be the treatment. The outcomes could potentially be related to policy-outcomes, but it would be case specific. The data-sets would be constructed using a legion of RAs.

   One example could be using COVID-19 and see how women political representatives managed the crisis. We could do it at the city level, and taking into account the gender of the mayor. We could do RDD comparing women who won by a small margin with places where women lost by a small margin. This assumes two candidates.

   For this, we will need to see how to obtain covariate balances between control and treatment. In any randomised design, to show that they are very similar. To do this, we can compare for the covariates we have by taking the average, making differences across groups, and see if differences are statistically significant. You can do this by running a probit/logit where dependent variable in the treatment and the covariates are jsut the covariates. Then you can just F-test that all covariates coefficients are equal to zero. We can also run a χ^2 test.

4. (**Theory**) **Discuss their theoretical modeling. Which aspects of the model are crucial to support the intuition behind their empirical results? Also discuss overall strengths and weaknesses of their model.** The principal differences between the Downsian and the citizen-representative model varies on the commitment, and it is one of the principal facts analysed by the empirical analysis.

5. (**Measurements**) **Are there any issues in their measurement of preferences? What are the key assumptions behind their measurement strategy? Discuss specific components of the data that justify their measurement strategy.**

   They use a citizen-candidate model in which the citizen wants to implement the policy she prefers, while in the traditional Downsian model with single-peaked preferences the candidate of the party only cares about winning. The authors put the paper in the house race between these two model, and put the findings on the side of the citizen-candidate model. On a note, the citizens candidate in the GP setting, specially in the case of women, seem to be realistic as people know each other in the village.

   Measurement of preferences is through requests, which could entail sombre problems. For instance, what is the coding used? What is the emphasis given? It is a black-box variable, ex-ante. Their measurement strategy relies on the fact that village level information is supposed to be reliable, because it is not provided by the Pradhan, and it is easy for the villagers to provide the information.

6. **(Identification) How does having two different states in the sample help identification?**

   Having geographical variation allows us to control for region fixed effects thus discarding some inherent trends that could confound the causal effect.

   On a side note, having two states also improves external validity.

7. **(Mechanisms) Are the mechanisms behind their empirical results crystal-clear? Discuss the analyses the authors conducted to get at the mechanisms in two dimensions: (1) key channels through which voters influence policy decisions; and (2) diverse ways in which female Pradhans differ from male Pradhans.**

   The paper precisely test out the mechanisms through which the independent variable is affecting the dependent variable. They single out one mechanism rejecting all the others.

   How are women different than male Pradhan's? What mechanisms make female Pradhan affect policy outcome wrt men?

   - Most of them are first-term politicians. Authors show, by comparing exogenous first term male Pradhans (thanks to the reservation rules), that this is not the mechanism.
   - Social interaction. But is is ruled out.
   - Women have higher socioeconomic backwardness. Their cost of running is higher. They show that this is not relevant by comparing with minorities, like scheduled casts and scheduled tribes.
   - The possibility to be reelected is lower. Without reservation, women know they are less likely to win. So using the data on the male pradhans that expect to be imposed a reservation, they compare with pradhans who are not going to be elected.
   - **Main mechanism behind the causal effect:Their own preferences are more aligned to the preferences of the people and thus serve them better.**

   The mechanism outlined by the authors is the following: *"Individual women are thus not particularly responsive to the needs of women and men in their communities. Rather, because their own preferences are more aligned to the preferences that they end up serving them better. This also alleviates the concern effect may be temporary because women are on their 'best behavior' they are conscious of being part of a social"*

4.4 Pande (2004) on Minorities

1. **(Modeling) This paper is closely related to Chattopadhyay and Duflo (2004, Econometrica) in the nature of the research subject. Discuss key similarities and differences in modeling decisions and the reasons for the differences.**

   The key similarity is that both take the same source of randomness, that of the GP mandate quotas. Also, the theoretical discussion is very similar, as both papers provide empirical evidence towards the citizen-reprsentative model.

   The main dissimalirity is with regards the independent variables of interest. Duflo includes an interaction term of the reservation indicator with the intensity of complaint differences and also average complaints, while Pande relies on a simple indicator for the reservation.

   In terms of theory, Duflo talks about the the provision of public goods and which goods are provided, while Pande looks at expenditure in different areas such as education, job training or transfers.

2. **(Identification) What are the key institutional features that the author exploits for the identification of causal effects? What are the key concerns regarding the identification strategy that the author tries to address?**

   That the setting of the quotas becomes more exogenous across time. In words of the author: *"My empirical analysis exploits the diachronic variation in the extent of political reservation enjoyed by a group in a state. The cause of such variation is defined by Section 3 of Article 332. Section 3 states that the proportion of jurisdictions reserved for scheduled castes (scheduled tribes) should equal, as nearly as possible, the population share of scheduled caste (scheduled tribe) in the state. Moreover, the only permissible basis for changes in the extent of reservation enjoyed by a group in a state is changes in the census estimates of the group's population share in that state."*

   To address these concerns, she controls for quite a bunch of stuff.

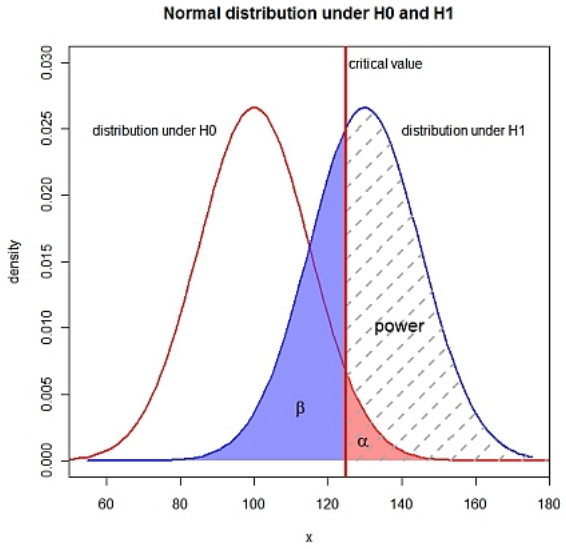
3. **(Results) Are the results perfectly convincing and fully explained? Discuss the results and weaknesses of the author's interpretations.**

   One of the striking results is that scheduled tribes invest less in education. Is because they are dumber? Author could have given more information about the budget and where the expenditure goes up.

   Also, why scheduled castes (SC) care more about job quotas than welfare transfers, while the contrary for (ST)? According to the author, this is due to the fact that relative to scheduled tribes, scheduled caste individuals are both more educated and geographically more dispersed.

   Moreover, population share comes from the census, so she does an interpolation from there, but there is nothing in the paper... BLACK BOX! Census share and actual population share? One thing that could have been done is to see the census share and actual population share to be plotted between treated and non treated. It is unclear which geographic portion of the data is givin the variation between these three variables to give identification of the reservation policy.

4. **(Additional Analyses) Are there additional analyses you could have done to gain deeper understanding of the mechanisms behind the results? Discuss.**



5 Fixed Effects Estimation and Difference-in-Differences

5.1 Evaluation of Giulia's presentation

- Structure of slides: Good. Some tables were too small to be read
- Comment about the paper: Redistricting (as in Pande's paper) creates an exogenous source of variation in districts that is very important for identification.
- Question handling and overall interaction with audience:

5.2 Handout on Fixed effects paper (2004)

5.2.1 Summary

We usually have:

$$y_{ist} = A_s + B_t + cX_{ist} + \beta I_{st} + \varepsilon_{ist}$$

where I_{st} is the intervention, which takes value 1 post treatment and 0 pretreatment.

Let's think of two problems: **bias and standard error**. For bias, the situation we might have is non parallel pre-trends. We can control for linear trends, in particular group specific linear trends. For this, we need data before treatment, and a fast treatment effect.

We can also 2SLS-FE. In Stata, the IV is IVreg. the fixed effects is xtreg. Then, 2SLS-FE has ivxtreg.

Now standard errors, the main problem today.

5.2.2 Type I and II and power of a test

Type I error, given by α, is the probability of rejecting H_0 when this is true. Type II error is given β, and given by accepting a false H_0 (or fail to reject H_0). The power of the test is $1 - \beta$, and it is the probability that you are not going to do a Type II error.

1. **Discuss (all) econometric issues that arise in the usage of difference-in-differences methods. Elaborate specific ways that we can detect each problem in the data.**

   The main econometric issue that arises is that of serially correlated outcomes and inconsistent standard errors, which leads to gross over-rejection of null hypothesis of no effect.

   To check whether our data has serially correlated outcomes, we can look at the correlogram of the outcome variable, by estimating the autocorrelation coefficients of the outcomes. Also, we can run several placebo

tests, by doing simulations of placebo treatments at random in our data. In this tests, we would expect to reject the null hypothesis of no effect $\beta = 0$ roughly 5 percent of the time when we use a threshold of 1.96 for the absolute t-statistic.

Another econometric issue that is not explicitly mentioned on the paper is that of parallel pretends. To see if this holds, the average outcome variable across different groups can be plotted against time to see if, previous to treatment, both trends are parallel or not.

Another note on Lim's class is that even if we do not have serial correlation in the outcome, we might have correlation between individuals in a given state i. To account for spatial correlation will decrease Type I error. We can also cluster.

Three features contribute to gross over-rejection of the null hypothesis:

(a) Typically, our dependent variable is something very serially correlated, like wage, (un)employment, health...

(b) Another thing is serial dependence in our independent variable, like the treatment.

(c) Finally, panel data usually have a long panel. Many variation comes from serial dependence, and long panels bad then.

2. **For each of the issues discussed above, discuss specific solutions that are employed and pros and cons of each solution.**

   To solve serial correlation, the first thing we might want to do is to put a parametric specification on the errors, for instance $\epsilon \sim AR(1)$. You can also collapse pre-post data, which solves overrejection but has also some problems, such as the fact that it has low power, so we have a high probability of commiting a Type II error[5]. A final solution is block-bootstrap, which imposes a block variance structure, by preserving the sample dependence using time by block. We want to take the entire time of the given state when we decide to bootstrap. Bootstraping is done by state but getting out a given observation across time. Bootstrap, however, does not work that well when we dont have that many states. Data should be large enough to be a good approximation of the population. If this data are not representative... bad.

   Finally, another solution would be flexible estimation of variance covariance, using white- glice method.

3. **Overview the Monte Carlo exercise discussed in this paper. What are the specific features of the typical context of the DD method that cause the serial correlation problem?**

   The Montecarlo exercise of the paper relies on using CPS data to see what is the rejection rate of the null hypothesis of no effect in different settings. For instance, for micro-data on female wages from CPS leads to a great over rejection. Clustering SE at the state-year level does not solve for it. Aggregating the data also does not work, but when we consider serially uncorrelated laws (treatment is applied only once every two times) the over-rejection disappears. Also, when a parametric structure is imposed in the model, with different levels of autocorrelation, in the case of an AR(1) only negative (or zero) autocorrelation provides sensible results.

   The typical features in the context of DD that cause serial correlation are:

   - Unwise choices when clustering SE
   - Serially correlated treatments (like laws over different periods).
   - Not aggregating the data correctly.
   - Unwise specification of the structure of errors autocorrelation.

4. **Discuss each of the solutions to the serial correlation problem examined in this paper.**

   - The first solution is to use block bootstrap, which is a variant of the bootstrap that maintains the autocorrelation structure by keeping all the observations that belong to the same group (e.g. state) together. However, this methods performs worse as the number of states declines.

[5]The initial problem was Type I error, but now it seems that with this method, we go to the other direction.

- The second solution is to ignore time series information, by averaging the data before and after the law and run the estimation in a panel of length 2. This method only works when the law is passed at the same time for all treated units.
- The third solution is to estimate an empirical variance-covariance matrix. If the autocorrelation process is the same across all states and there is no cross-sectional heteroskedasticity, the data are sorted by states and years, and the variance-covariance matrix of the error term is block diagonal. We can use the variation across the 50 states to estimate each element of the matrix, and use this matrix to compute standard errors. Under no heterkedasticity, this method will produce consistent estimates of the standard error asymptotically.

5.3 Handout on Snyder and Stromberg (2010)

1. (**Motivations**) **Discuss the way that the authors motivate the importance of their research questions, focusing on the broader variation in media environments.**

   Having a press that actively covers politics is essential for democracy. And also, empirical evidence of the effects of active media coverage was scarce in the literature, prior to the paper at least. This is because media coverage is endogenous to most outcome variables of interest.

   To overcome this problem, the authors rely on the fact that the economic geography factors that determine media markets are generally quite different from the political geography factors that determine congressional district boundaries. This is defined as the congruence between newspaper markets and congressional districts.

2. (**Identification**) **What are the key identification problems that would arise if we regress outcome variables on the amount of media coverage? How does the authors' address the identification problems?**

   The authors address this issue by the exogenous variation that noncongruent areas between media coverage and districts create. Voters in noncongruent areas are exposed to less news about their representatives, since the newspapers sold in their area simply devote less coverage to their House representative.

   A key identifying assumption in the empirical investigation is that Congruence is not directly related to variables such as voters' intrinsic interest in politics.

3. (**Strengths and Weaknesses**) **Discuss key strengths and weaknesses of the paper. Are there any additional analyses you would have done?**

   Answer

4. (**Mechanisms**) **Discuss each of the mechanisms of the media effects that the authors establish.**

   According to the authors, the main force behind the results in this paper is that the number of articles that a newspaper writes about a House representative is increasing in the estimated fraction of the newspaper's readers who live in the associated congressional district, the Reader share.

6 Weak Identification and Altonji-Elder-Taber Test

We started the course talking about IV and RCTs. As motivation, we talked about potential outcome framework. Then, we got into the Moffitt RFT paper, related to Duflo's, but in high scale. Then we had an application in the Gerber paper. In Week 3 we had natural experiments, with Duflo and Pande's papers. Within these papers, we dealt with the Downsian versus citizen-candidate models.[6] Now, in Week 5, we have a similar discussion, using RDD techniques, on Downsian versus

1. **Subsample**. They play around with different subsamples. In the US, they have K-12, which are the grades from kindergarden to 12th grade. A particular way to chop is is between primary eight and then high school. You can also chop it out in 6, 3 and 3. In any case, the end of eight grade is a final stage before high school. The C8 sample are kids from 6 to 14. Some catholic C8 went to catholic high school, and some went to public high school. But most of public school go then to private high school.

[6] In Week 4, we had a discussion on the DiD setting and the Stromberg paper

Authors note that people going to public or catholic high school are pretty similar if they both went to catholic C8. However, for some of the key variables the C8 kids that went to public HS differ from the ones that went to catholic HS. Therefore, it sucks.

2. **Modeling and Identification**.

   In this paper they are trying hard to get selection on unobservables and selection on observables. The key is going to be the selection equation. We want to understand the choice of catholic high school, using a probit:

   $$CH = 1(X'\beta + \mu > 0)$$

   where μ follows a standard normal distribution. And then other outcomes, such as HS college follows

   $$Y = 1(X'\gamma + \alpha CH + \varepsilon > 0)$$

   where $\epsilon \sim N(0, 1)$. What matters is the relationship between μ and ϵ, which causes the problems of (lack of) identification. This is a special version of the Roy model. We will estimate the joint structure of μ and ε as a normal with mean zero, variance 1 for each and covariance ρ.

   Question: what does semi-parametric and fully parametric specification, what does that mean? For instance, if you impose a structure in the error, you are estimating a fully parametric model. A semi-parametric specification relies on imposing a different structure $\mu = \beta + \mu^*$ and $\varepsilon = \beta + \varepsilon^*$, where β follows a normal. And now, you can relate Heckman two-step estimator with selection bias and explicitly derive the selection bias and control for it[7].

   The authors know that we cannot fully trust the normal distribution assumption. We cannot find a credible IV that satisfies exclusion restriction, so we will rely on bounds on the causal effect.

3. **Modeling and Identification (2)**.

4. **Key Assumptions of the Main Model**.

   The upper bound is coming from the OLS. In the OLS we assume there is no self-selection. The question boils down to the lower bound of the causal effect.

   For this, first let us decompose the outcome variable:

   $$Y = 1(Y^* > 0)$$

   The causal effect is given by the causal effect and the rest:

   $$Y^* = \alpha CH + N'\Gamma = \alpha CH + X'\Gamma_x + \xi$$

   where Γ_x is the observed heteroegeneity.

   Therefore, we have,

   $$Y^* = \alpha CH + X'\gamma + \varepsilon$$

   where we have $cov(x, \varepsilon) = 0$. Now, run a regression of CH on $X'\gamma$ and ε. This yields two coefficients, $\phi_{x'\gamma}$ and ϕ_ϵ. We have that these two coefficients are a measure of selection on observables and unobservables, respectively.

   The upper bound is pinned down by condition 2, which argues $\phi\epsilon = 0$, and the lower bound is pinned down by condition 1, which states $\phi_\varepsilon = \phi_{x'\gamma}$.

5. **Restatement of the key assumption**. Recall the key assumption are that observables were selected quite randomly. Also, we need that there is a lot of observables and that the ones we have do not drive the distribution. In general, this does not happen as typically selection on observables is higher than selection on unobservables, as researchers pick the maximum possible observables that might drive the effect. Therefore, we have a condition which is $0 \leq \phi_\varepsilon \leq \varepsilon_{x\gamma}$, which states that observables typically will be more powerful predictors than unobservables[8].

[7] Check a paper called "Empirical Content of the Roy Model"

[8] This is the most usual sitaution, if your data are decent enough.

Finally, condition 4 is used to quantify the ratio of selection on unobservables to selection on observable, to actually negate all the positively estimated causal effect of catholic high-school. This will be the basis of the qualification exercise of the paper. It is a rewritten form of condition 1.

$$\frac{E(\varepsilon|CH=1)-E(\varepsilon|CH=0)}{Var(\varepsilon)}=\frac{E(x'\gamma|CH=1)-E(x'\gamma|CH=0)}{Var(x'\gamma)}$$

The derivation from condition 1 takes the definition of the estimators:

$$\frac{Cov(\varepsilon,CH^*)}{Var(\varepsilon)}=\frac{Cov(x'\gamma,CH^*)}{Var(x'\gamma)}$$

But then, divide both sides by $Var(CH)$, and the condition 4 is obtained.

Now, note that since $CH=X'\beta+\tilde{CH}$,

$$Y^*=\alpha\tilde{CH}+\alpha X'\beta+X'\gamma+\varepsilon=\alpha\tilde{CH}+X'(\gamma+\alpha\beta)+\varepsilon$$

Therefore, we have

$$\text{plim }\hat{\alpha}=\alpha+\frac{Cov(CH,\varepsilon)}{Var(\tilde{CH})}+\frac{Cov(X,\varepsilon)}{Var(x)}=\underbrace{\alpha}=0+\frac{Var(CH)[E(\varepsilon|CH=1)-E(\varepsilon|CH=0)]}{Var(\tilde{CH})}$$

and done.

7 Regression Discontinuity Design

7.1 Lee and Moretti (2005): Do voters elect or affect policies?

1. **Motivation**. The angle of the research question is pretty different from Duflo's and Pande's papers. The authors are interested in random assignment in electoral strength on the voting behaviour of congressmen in the US house. To do this, they consider electoral strenght in candidates that won by a tight margin. The thing is that these districts are pretty comparable in terms of voters, but the electoral strenght of the candidate is as random. Thus, the observable characteristics of districts with a tight Republican are very similar to districts with a tight Democrat win.

   Also, the difference is between full divergence and partial convergence in this paper. This is more difficult to test than Duflo and Pande's paper questions, which test full convergence vs partial convergence.

2. **Identification**. The main assignment the guys are interest in is the roll-call at time $t+1$ where x axis is the democratic voting share, and we have a high total effect of elect and affect.

   Problem is incumbent advantage.

3. **Interpretation of the Results**.

4. **Alternative Strategy**. We could try structural estimation.

5. **Context/External Validity**. See recording of last 45 minute of lecture 5. i am tired

7.2 Ferreira and Gyourko (2009). Political partisanship

1. Both are RD papers but random assignment differs. Random assignment

2. The most important contribution of the paper is the exploration of the mechanisms

3. Because they take elections where mayors are elected directly by the populations. However, the sample is 5% of US cities but are big cities and thus encompass a lot of population.

4. Population homogeneity, Tibout competition (provision for public goods) and strategic extremism through lack of newspapers. Using regressions.

8 Structural Estimation - 1955 paper

We have three reasons to use structural estimation:

1. Addressing endogeneity by modelling the choice process.
2. Welfare measurement.
3. To run counter-factual policy simulations.

In the paper, the author talks about useful and necessary information. Gaining theoretical information of the choice process is very important. What information is necessary depends on the objective of the decision makers. To illustrate this, the author proposes three scenarios.

- In the first scenario, the firm is the agent.
- In the second scenario, the agent is the government. This government might want to maximise tax revenue θq
- In the third scenario, the agent is the government again, but it wants to maximise the output given that taxes satisfy a given constraint $T = T^*$.

The way author defines useful information is the minimum set of information that helps us to decide how to make the optimal decision given the objective.

In the first scenario, useful information are the structural parameters such as α and β and the taxes θ. For the point of the government, θ is not useful information, because it is chosen by the government. However, the author shows the optimal tax is $T^* = \alpha/2$, so only α is useful information.

This exercise sheds light into the idea that sometimes we don't have to have knowledge about all parameters, we don't need to identify them all.

What are useful info in the third setting? α is. Then we have that $T^* = \theta(\alpha - \theta)(2\beta)^{-1}$ so useful information is also β. θ is chosen.

Now we deal about θ. We can think that it didnt change, that it didn't change but can change, and that it actually changed.

9 Structural Estimation of Dynamic Models

9.1 Optimal Replacement of GMC Bus Engines: Rust (1987)

This paper formulates a regenerative optimal stopping model of bus engine replacement to describe the behavior of Harold Zurcher, superintendent of maintenance at Madison Metropolitan Bus Company. This is done to provide a simple, concrete framework to illustrate two ideas: (i) a "bottom-up" approach for modelling replacement investment and (ii) a "nested fixed point" algorithm for estimating dynamic programming models of discrete choice.

The optimal stopping rule is the solution to a stochastic dynamic programming problem that formalises the trade-off versus minimiing maintenance costs versus minimizing unexpected engine failures.

1. **Summarise the key contributions of the paper in the following two dimensions: (1) understanding of investment; and (2) econometric modeling of optimal stopping problems.**

   The first contribution is to illustrate the "bottom-up" approach that uses a micro-theoretic model to derivate aggregate replacement investment from individual optimizing behavior. Investment is modeled as a regenerative stochastic process, where a regeneration corresponds to replacing an existing used asset with a new one.

   The second contribution is to illustrate a new estimation method that allows to compute maximum likelihood estimates of the primitive parameters of a class of controlled stochastic processes, even though there is no analytic formula for the associated likelihood function. In particular, the model shows how one derives the sample likelihood function for the regenerative process as the solution to a regenerative optimal stopping problem.

2. **What is the main point of dynamics in this paper? Elaborate the main features (payoffs and costs) that necessitated a dynamic model.**

   Dynamics account for the tradeoff between making normal maintenance on a bus now and use it further with the fact that in the future it may not be useful or scrap it altogether, put a new engine, and use the parts to minimise costs of future maintenances.

   The main elements are the vaue function that accounts for the maximised futured discount utility of making maintenance on a bus or scrapping it and using for future repairs. Then, there is an optimal stopping strategy given by i_t that depends on the mileage x_t and operating expenses.

3. **Reconstruct the model with exponential distribution. Discuss the key issues of this model**. Answer

4. **Reconstruct the main model**

   - **Reconstruct the key structural components of the model**. Answer
   - **Discuss the estimation procedure. what are the key assumptions in the model that simplify value function calculation and choice probabilities? Formulate value functions and choice probabilities as a function of the parameters of unobservable components of the utility functions.**
     Answer

9.2 A Political Economy Model of Congressional Careers (Diermeier et al (2005)

The main goal is to quantity the returns to a career in the United States Congress, using a dynamic model of career decisions. With this model, reelection probabilities are estimated, as well as the effect of congressional experience on private and public sector wages. Then, the value of a congressional seat is estimated. Finally, it is assessed how an increase in the congressional wage or the imposition of term limits would affect the career decisions of politicians and the returns from a career in Congress.

Thus this paper answers the question of what are the returns to an individual from a career in politics? Also, what about imposing term limits to make politicans accountable?

The framework of the paper allows to sort out the relative importance of (i) utility politicians derive from being in office and (ii) the monetary returns to a career in Congress.

Previous papers have estimated choice models that do not take into account the dynamic aspects of politicians' career choices. A second overseen problem in existing studies is that they ignore the career prospects of politicians after they leave congress. Third, selection bias created by politicians' decisions about whether to run for reelection are ignored. A fourth problem is unobserved heterogeneity of politicians and their effect on career choices.

The main findings are that congressional experiences significantly increases post-congressional wages, but these diminish with additional experience. Also, nonpecuniary rewards from being in Congress are rather large. In particular, monetary returns alone cannot explain the observed behavior of politicians and the effect of the congressional wage on their behavior is quite small. Third, politicians' unobserved skill play an important role throughout their congressional careers. Selectivity bias is found to be modest. Finally, it is found that the imposition of term limits would substantially increase early voluntary exit from Congress and significantly reduce the value of a congressional seat.

1. **Summarize the key contributions. Also, discuss the primary reasons they estimated a dynamic structural model. Why did they need a structural approach? What is the role of dynamics?**

   The main contributions of the model os to incorporate the dynamic traits previously overlooked in the literature, as well as incorporating the career prospects of politicians after they leave congress in their decisions. Also, selection bias and unobserved heterogeneity are addressed.

   The structural approach is necessary to estimate the dynamic model. Dynamics accounts for previously uncovered facts related to dynamic decisions such as whether to remain in congress or not or whether be a Congressperson or not.

2. **Reconstruct the estimation procedure used in this paper (How would you structure the program for estimation?**

   First, wage is estimated using a Mincerian regression on different observables related to human capital.

3. **Reconstruct the procedure for counterfactual experiment on the wage increase (How would you structure the program for simulation)?** Answer

4. **Discuss identification of structural parameter: non-pecuniary value of the seat, unobserved type ("skill", "achieve") of politicians.**

10 Non-AR(1) Markov Chain

(a) well

(b) well

(c) We have to compute the invariant distribution analytically in terms of γ and π. This is the solution to the problem

$$\boldsymbol{f}\boldsymbol{\Pi} = \boldsymbol{f} \tag{1}$$

where, noting that in the initial distribution $\pi_i = \pi = 0.2$

$$\boldsymbol{\Pi} = \begin{pmatrix} \gamma + (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi \\ (1-\gamma)\pi & \gamma + (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi \\ (1-\gamma)\pi & (1-\gamma)\pi & \gamma + (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi \\ (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & \gamma + (1-\gamma)\pi & (1-\gamma)\pi \\ (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & (1-\gamma)\pi & \gamma + (1-\gamma)\pi \end{pmatrix}$$

and $\boldsymbol{f} = [f_1, f_2, f_3, f_4, f_5]$

First of all, denote $\pi + (1-\gamma)\pi = a$ and $b = (1-\gamma)\pi$

Then we have:

$$\boldsymbol{\Pi} = \begin{pmatrix} a & b & b & b & b \\ b & a & b & b & b \\ b & b & a & b & b \\ b & b & b & a & b \\ b & b & b & b & a \end{pmatrix}$$

The problem from 1 becomes:

$$\begin{aligned} af_1 + b(f_2 + f_3 + f_4 + f_5) &= f_1 \\ af_2 + b(f_1 + f_3 + f_4 + f_5) &= f_2 \\ af_3 + b(f_1 + f_2 + f_4 + f_5) &= f_3 \\ af_4 + b(f_1 + f_2 + f_3 + f_5) &= f_4 \\ af_5 + b(f_1 + f_2 + f_3 + f_4) &= f_5 \end{aligned}$$

And it follows that by symmetry of the equations that $f_i = f = \pi$ for $i = 1, \ldots, 5$.