PS0700
Research Designs in Political Science: Quasi-Experiments, Natural Experiments, and Passive Observational Designs

Political Science Research Methods

Professor Steven Finkel

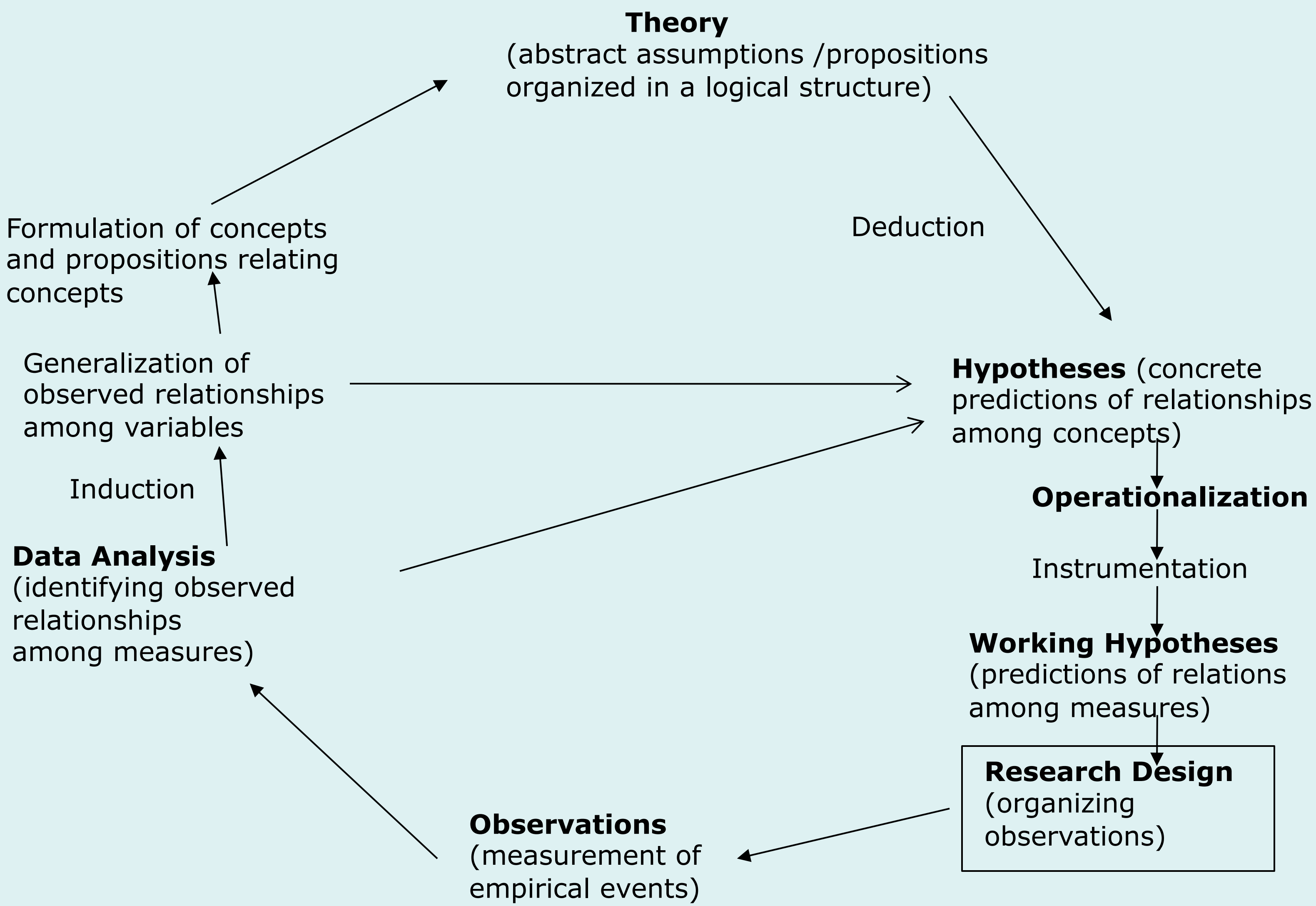
Fall Semester 2022

Week 5

Goals for the Sessions

- Discuss different kinds of designs that fall short of the experimental ideal of random assignment and researcher-controlled manipulation of the independent variable
- Discuss strengths and weaknesses of "quasi-experiments", "natural experiments", and "passive observation" designs
- Show the relevance of these designs to contemporary political science research

A Model of the Research Process



"Research Designs involve setting up a research project so that research questions can be answered as unambiguously as possible. The objective of a good research design is to establish causal relationships and to assess their generalizability."

Meier, Brudney, and Bohte, *Applied Statistics for Public and Non-Profit Administation*, p.52.

What do you do when a Classical Experimental research design is not possible?

- In some instances, the researcher still has the option of controlling some kind of manipulation of the independent variable and administering it -- non-randomly -- to some units and withholding it from others. This is called a **"quasi-experiment"**
- In other instances, some units are subjected to a random or near-random manipulation by "nature," or forces outside the researcher's control, and the researcher can compare those units on the dependent variable. This is called a **"natural experiment"**

- In yet other instances, the researcher simply observes units that have not been explicitly "manipulated", but that nevertheless have different levels of the independent variable, and the researcher compares their responses on the dependent variable. This is called a **"passive observation"** study, of which there are many different kinds , depending on how many units there are and how many observation points there are for each unit
- The terminology can be confusing, as many books use different definitions and even practicing social scientists get them mixed-up. For example, the natural experiment could possibly be characterized as a kind of "passive observation" design because the *researcher* didn't actively manipulate anything, nature did, and the researcher just "observes" the results.
- The important issues are the **random or non-random** nature of a manipulation, and whether a manipulation it is done **actively**, by **"nature",** or **passively observed** by the researcher

- A corollary to this statement is that **all** of these designs have certain limitations, owing to the lack of **pure** randomization and/or the lack of the kind of **researcher control** over the treatment (X)
- Different non-experimental designs, though, still are associated with different levels of confidence in estimating causal relationships ("internal validity") and in generalizing the results ("external validity"). As a general rule, we prefer designs with
  - treatment and control groups that are "closer" to being randomly assigned
  - pre-tests
  - more time points
  - more units
  - more researcher control over the introduction of the independent variable ("treatment")
  - more levels of analysis (e.g., "macro" and "micro" level)
- The better designs, however, are not always possible to implement, so we need to know how to maximize the power of lesser designs!!!

The Quasi-Experiment

- Classic Definition from Cook and Campbell 1979:

  "Experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred. Instead, the comparisons depend on non-equivalent groups that differ from each other in many ways other than the presence of the treatment whose effects are being tested."

The Classic Quasi-Experiment

	Pre-Test		Post-Test	Difference
Treatment Group	M_{1t}	X	M_{2t}	M_{2t} - M_{1t}
Control Group	M_{1c}		M_{2c}	M_{2c} - M_{1c}

Critical Features:

- **Non-random or voluntary assignment to treatment or control group**
- Sometimes: less control over the form and the level of X, i.e., the treatment
- Pre-test to measure baseline outcomes and subsequent change
- As in classic experiment, the “difference in difference” is the estimation of the causal effect of X

The QE design is the workhorse of much applied policy and evaluation research!!

- Microfinance and other international development projects
- Job training and welfare to work programs
- Educational reforms
- Effectiveness of civic education programs
- Anti-delinquency programs
- Anti-drunk driving programs

Examples

- "Bolsa Familia" Anti-poverty program in Brazil
  – Conditional cash transfers to eligible poor families in Brazil. Compare to control families: poor families not enrolled in the program
  – Attempted in NYC through a randomized trial with mixed results
  – **Why QE? 1) Treatment exposure was pegged to eligibility criteria, not randomly assigned; 2) Intervention was directly imposed by the state (the "researcher")**

- Finkel, Kenya Civic Education Study (2001-2002)
  – 1000 Kenyans interviewed as they entered one of 161 civic education workshops conducted in February-May 2002 in run-up to national elections; 1000 "control" individuals from same neighborhoods interviewed simultaneously. Individuals re-interviewed in December, just before elections take place. [Results suggest that civic education has positive effects on many democratic orientations such as tolerance]
  – **Why QE? 1) Treatment exposure was based on self-selection, not random assignment; 2) Intervention was directly imposed by the NGOs conducting the civic education**

Strengths of Quasi-Experiments with Pre-Tests

- QE studies with pre-tests can control for the baseline levels of Y for all groups – the *starting points* on the dependent variable
- The reseacher directly observes *changes* in Y, and (usually) knows that X came before the changes in Y were observed
- QE studies often have relatively high *external validity*, as they involve real world observations, without laboratory or other constraints imposed through classic experimental methods

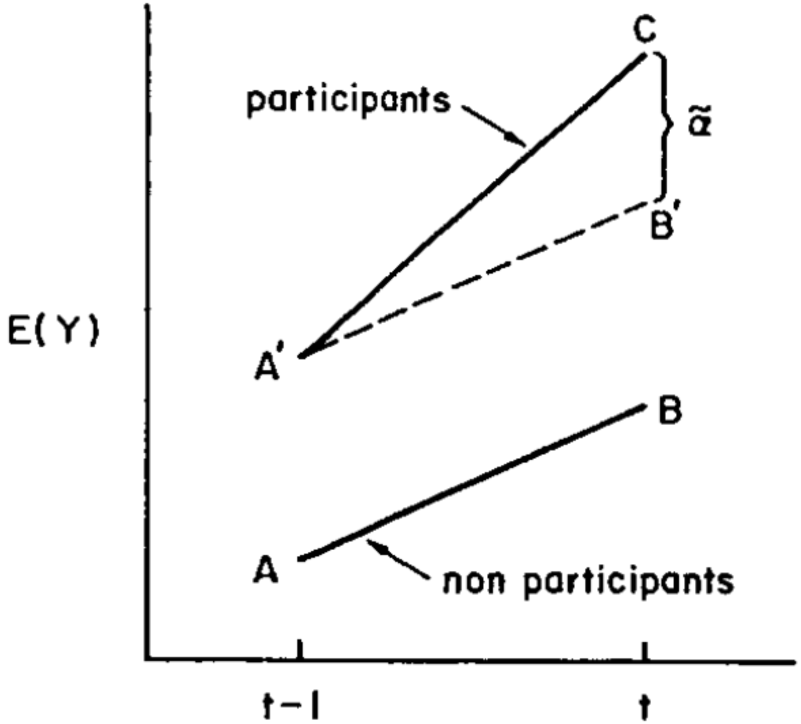
Potential Problems in Estimating Causal Effects with QE Designs

- As with the classic experiment, the QE researcher may not know the true length of time it takes for X to cause Y, so the design may not capture it
- As with field experiments, there may be problems in implementing the QE design
  - possible non-compliance and drop-out
  - possible unknown variation in treatment
  - possible "contamination" of the treatment and control groups through spillover effects
  - possibly unrepresentative sample sites

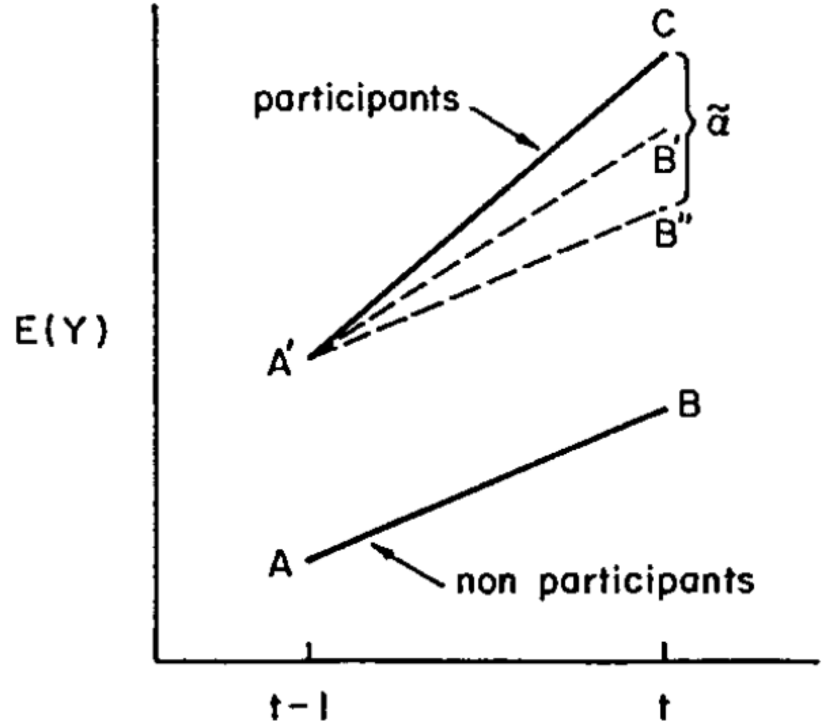
Biggest Problem with QE studies: “Selection Bias”

- Because the treatment group differs from the control group on a lot of possible Z factors, we can‘t *completely* rule out the possibility that Z confounds the process and that, therefore, there is not really a causal effect of X on Y
- There is often *self-selection* into the treatment group, and the kinds of things (Z) that lead individuals or units to select themselves into the treatment may also influence changes in Y over time (as in the Finkel study from slide 11)
- The presence of a pre-test is *very* helpful in ruling out some kinds of selection bias
- But there are still problems with selection bias, all of them rooted in the non-random assignment to treatment and control
- This means that the control group may not be a proper “counterfactual” group to compare with the treatment group

- **“Selection – History”** interaction: the treatment group differed from the control group, and this difference led to a reaction to something else aside from X that happened after the pre-test
- **“Selection – Maturation”** interaction: the treatment group was already changing on Y at a faster pace than the control group
- **“Selection – Regression”** effect: If people who are generally very low (or high) on Y are selected for treatment, their gains on Y will be disproportionately positive (or negative) compared to the control group even in the absence of a treatment effect
- **Another way to express this problem: the assumption needed to justify QE designs as showing a causal effect of a treatment is that the treatment group, had they *not* been treated, would have had the same trend in the outcome from pre-test to post-test as did the control group – this is known as the “parallel trends” assumption**



Parallel Trends Assumption Holds



Parallel Trends Assumption Does Not Hold

- Panel (a) is fine. In panel (b), though, the treatment group would have changed more than the control group, even in the absence of treatment, so QE overestimates the treatment effect as (C-B") instead of (C-B')
- Always ask: How were the groups different before X was introduced, and how could those differences have accounted for the differences in the results/trends observed for the two groups?

How to Overcome these Potential Threats?

- "Matching" of the treatment and control groups on as many Z variables as possible *before* the treatment is introduced, so that the groups are as "balanced" as possible in a non-randomized design. [Problems: we need to know the relevant Zs and measure them; and this does not control for possible "unmeasured" Zs]
- Statistical controls for Zs in the data analysis phase. See if X still leads to changes in Y once the Zs are controlled. We will learn some of this when we talk about "multivariate analysis" later in the class
- Add more waves of observation – easier to test "parallel trends" assumption and deal with it statistically
- In policy research, use the "oversubscription" or "phased roll-out" methods of implementation of treatment to ensure control groups that are as similar as possible to the treatment groups on both observables and unobservables. (This could lead to a "true" experimental design too!)

Quasi-Experiments without a Pre-Test:

- A less powerful version of the QE has the same general features *but no pre-test*. That is, some stimulus (independent variable) is administered to a non-randomly-assigned treatment group and the levels of the dependent variable are compared to a non-randomly-assigned control group.
- It looks like this:

Quasi-Experiments without a Pre-Test: The Non-Equivalent Control Group Design

		Post-Test
Treatment Group	X	M_t
Control Group		M_c

- Example: Finkel's *Civic Education and the Mobilization of Participation in New Democracies* (*Journal of Politics*, 2002): estimate the effect of exposure to civic education workshops in Dominican Republic and South Africa by comparing participants with "matched" non-participants at one point in time, after the workshops had taken place

Problems with Causal Inference?

- The lack of randomization and lack of pre-test are crucial, because:
  - it is difficult to rule out the possibility that the treatment group differs from the control group on a whole lot of Zs, some of which may determine Y as well ("spuriousness problem")
  - A pre-test could at least control for differences in the *starting point* of the two groups. But with this design it is difficult to rule out the possibility that treated individuals were *already* more participatory before they were exposed to X (in the civic education case, before they attended the workshops)

- It may even be the case that individuals attended the workshops precisely because they were already participatory, in which case Y→X, not X→Y (“time precedence” problem)

- Sometimes we use the measurement process to find out how much X a person/unit was exposed to, and since this takes place afterwards, there may be more inaccuracies in our recording of this information than in the QE design with pre-test. Did the unit really get exposed to X or not?

A “Near” Experimental Design: The “Natural” Experiment

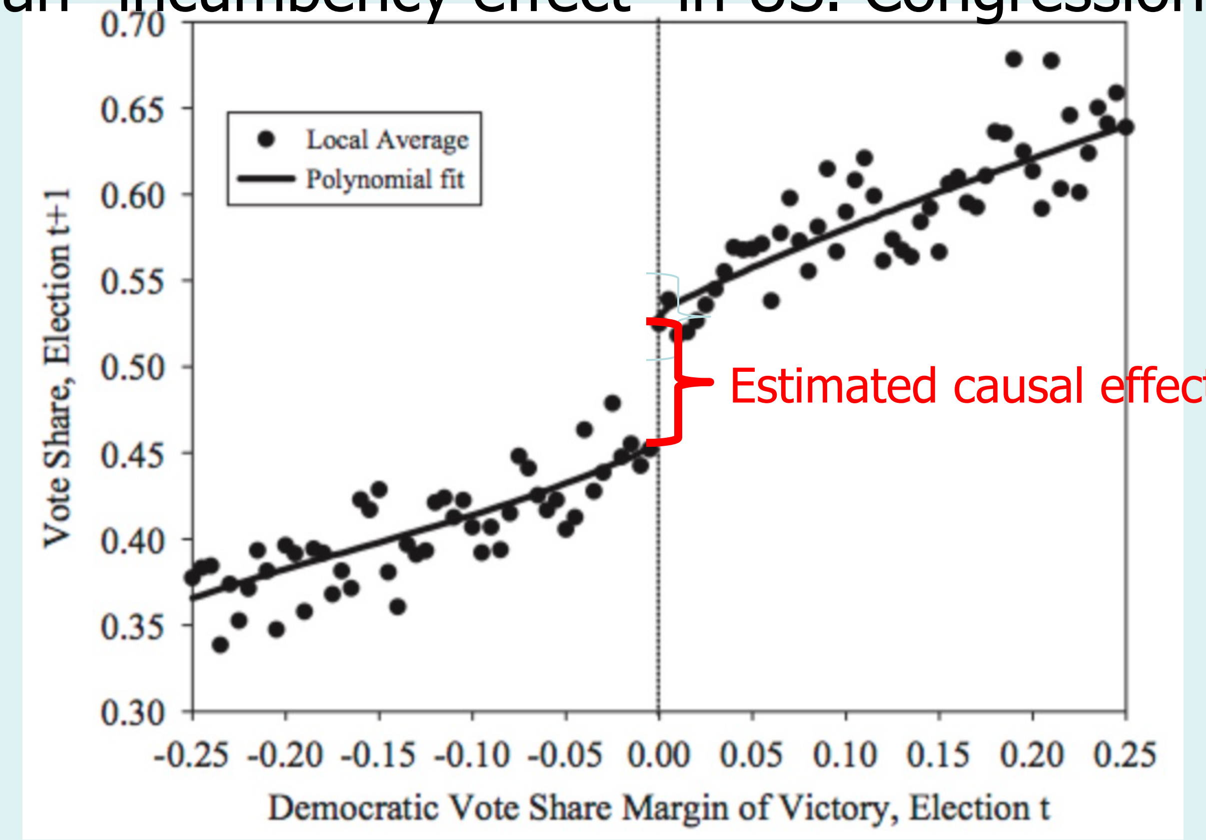
- “Natural experiments” are becoming more popular in political science. They have two distinguishing features:
  - They have active manipulation of a treatment X, but the manipulation is not done by the researcher, rather it is done by “nature”
  - Nature ”assigns” units to receive the treatment as close to randomly as is possible
- For example, during the Vietnam War males born in certain years were “randomly assigned” to be drafted based on a lottery that gave high or low numbers depending on the month and day of birth (e.g., May 15 could have been #1, December 2 could have been #2, etc.).
- Researchers then used this assignment to see the effects of military service on political attitudes, income, or whatever the interests of the researcher may have been (Erikson and Stoker, *APSR*, 2012)

- The more "random" the natural manipulation, the more the design resembles a true experiment. Scholars would like to be able to claim that the manipulation by nature was "as if" random; to the extent that this is plausible, it gives the design more power
- Examples from Class Web Site: Posner, "The Political Salience of Cultural Difference (*APSR* 2004)
- Classic Example from Economics: What is the effect of education on later income? Causal inference difficult because of selection biases – the kinds of people who typically choose higher education may be different in abilities. So economists use "quarter of birth" as a randomized push to obtain education, since people born in the last quarter of the year stay in school longer than people born in the first quarter of the year (due to mandatory schooling until age 16, first quarter kids can drop out earlier than last quarter kids). Comparing differences in earnings among people from different birth quarters is a "natural experiment". See Angrist and Krueger, *Quarterly Journal of Economics* (1991).

Regression Discontinuity Designs (RDD)

- Another kind of "near random treatment assignment" study treatment is found in the "regression discontinuity design"
- In this design, there is a threshold or "cut-point" where units either get the "treatment" or don't
- The logic of this design is that units which are very close to the cut-point more or less randomly get across the threshold or don't, and that we can therefore compare those units to get the causal effect of a treatment
- For example, some anti-poverty programs give assistance only if families have household incomes below, say, $20,000 dollars a year. In the RDD design we could compare families with, e.g., $19-20 thousand dollars to familes with $20-21 thousand dollars of income; they are very nearly identical but those under 20 got the government assistance and those over 20 didn't.
- Whenever receiving a treatment depends on some fixed cut-point, can use the cut-point as a kind of exogenous factor that (almost randomly) determines "treatment" near the cut-point

Is there an “incumbency effect” in US. Congressional Elections?



- Note “discontinuity” of the line right at the cut-point of eligibility – this is the estimated “impact” of treatment (red bracket = causal effect)
- In this case the “cut-point” in the margin of victory of a Democratic candidate in an election. Cases very near the cut-point were really close elections at time t. Then you can see that the effect of winning at time t+1 for a Democratic is about 11 percentage points. Lee,

Pure Observational Studies

- A final class of designs are characterized by **BOTH** 1) the lack of active manipulation of a treatment X **either** by the researcher or by nature; and 2) the lack of random or "near random" assignment to treatment. These designs are called "pure observational designs", or "passive observational designs"
- **The researcher passively observes non-randomly assigned groups who have different levels of X, and who "received" those different levels of X in ways that the researcher does not control**
- These designs are the **weakest** in terms ''proving'' causality. They suffer from many of the potential biases we've talked about: Z confounding variables, differences between treatment and control group at baseline, etc.
- Nevertheless: causal claims are still possible with these designs even if they are harder to completely defend. And these designs are *very* common in political science. And there are *many* different kinds of these designs, some better than others.

Passive Observation or Single Shot Ex-Post Facto Design

		Post-Test
Units with Value (A) on X (“Treatment Group”)	X	M_t
Units with Value (B) on X (“Control Group”)		M_c

- Example: Our old friend, the hypothesis "Does the individual's financial conditions influence their vote?"
  - "Treatment" Group: Individuals whose financial situation got worse
  - "Control" Group: Individuals whose financial situation did not get worse
  - Single-shot observation: We conduct a representative survey of US adults, and we ask: Did people in our "Treatment Group" and our "Control Group" differ in their likelihood of voting for the incumbent party or opposition party candidates?
- Example: Our other old friend, the hypothesis "Does social trust lead to democracy?"
  - "Treatment" Group: Countries with high levels of social trust
  - "Control" Group: Countries with low levels of social trust
  - Single-shot observation: We use existing surveys around the world and divide countries according to their aggregate level of trust: Are "high" countries more likely to be democratic than "low" countries?

Strengths and Weakness of the Single-Shot ("Cross-Sectional") Passive Observation Design

- Strengths
  - *External validity* can be *really* high if the study has a good representative sample. We really see how X relates to Y in the real world, in the population at large
  - When manipulation of X is not possible, ethical, practical, or feasible, this design is often all that the researcher can do
- Weaknesses
  - Lack of randomization and lack of pre-test mean that the groups differ on a lot of Zs, some of which may determine Y ("spuriousness problem"). No pre-test means we can't rule out pre-existing differences that caused different starting points
  - We don't control X, so it may be that it came after Y and not before, or it may be that Y actually *causes* X and not the reverse
  - ALWAYS ASK: HOW ARE THE GROUPS DIFFERENT ASIDE FROM THEIR DIFFERENCES ON X, AND HOW COULD THOSE DIFFERENCES ACCOUNT FOR THE RESULTS THAT ARE FOUND?

How to Overcome these Potential Threats?

- Statistical controls for Zs in the data analysis phase. See if X still leads to changes in Y once the Zs are controlled. We will learn some of this when we talk about "multivariate analysis" later in the class
- But this cannot control for confounding due to unobservables, which, by definition, the researcher has not measured or included in the study!
- See if it is possible to induce – or have "nature" induce – changes in X via random processes, and then you have something closer to an experiment. Example: randomly encourage respondents in a pre-test survey to attend a civic education workshop, then look at "encouraged" and "non-encouraged" groups' changes over time

Alternative Method: Add More Time Points or Periods of Observations!!

- Add more observations on each unit so you can see how the dependent variable changes *after* an independent variable changes, and you can see more clearly the longer-term trends for groups that do and do not change on X over time
- When you have only a few time points and many units, it is called a *panel* study
- When you have one or only a few units with *many* time points, it is called a *time series* study

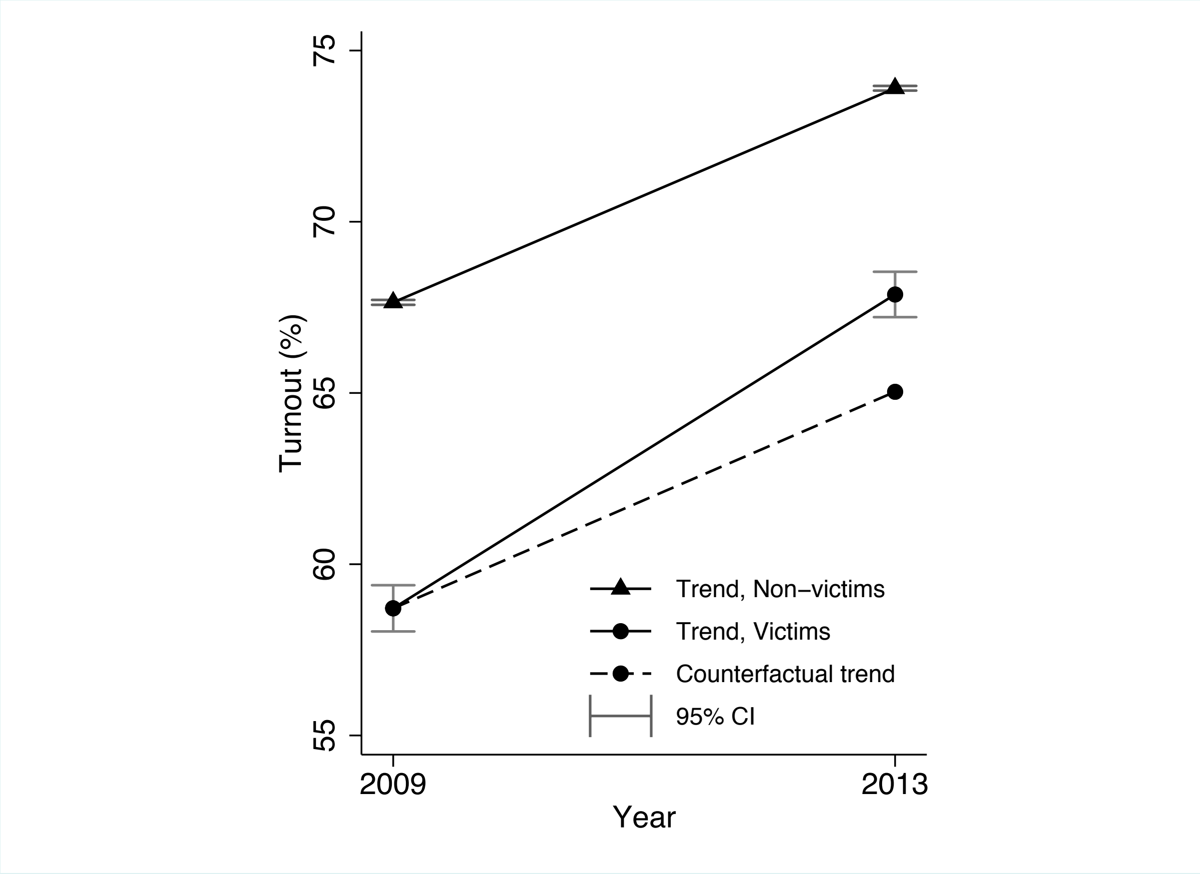
Panel Studies

- Example: Finkel (1985) on syllabus. Survey of US adults 1972-74-76 (a 3 "wave" panel). Ask about participation in elections and questions relating to political efficacy in all waves. Do people who participate in politics *change* in their efficacy over time more than people who don't participate?
  - Treatment group: People who participate at one point in time
  - Control group: People who don't
  - Dependent variable: Change in efficacy from one point to the next
- Advantages:
  - Can analyze *change* in dependent variables directly
  - Can use previous values of Y as "pre-tests" to control for Zs
  - With enough waves, can take into account Y→X as well as X→Y
  - Can add more Zs into the analysis statistically with matching or other methods as well

- Panel designs are the closest **observational** design to an experiment that we have in political science!!!!
- If you have two time points, then the "passive observational" panel design has exactly the same form as the Classic Quasi-Experimental Design on slide 9. The only difference is whether the "treatment" X is observed passively or whether it is actively manipulated by the researcher. The more time points pre-and post treatment, the better (this is true for QE and panel studies both!)
- But pay attention to "parallel trends" assumption and whether it is likely to be violated!
- Example from Class Web Site (recitation discussion)
  - "Why Demonstrating is Good for Kids", New York Times 3-12-2018.

Sønderskov *et al.*, "Crime Victimization Increases Turnout" *British Journal of Political Science* (2021)

Difference-in-Differences Observational Design



Treatment: Being a Crime Victim between 2009 and 2013 Elections

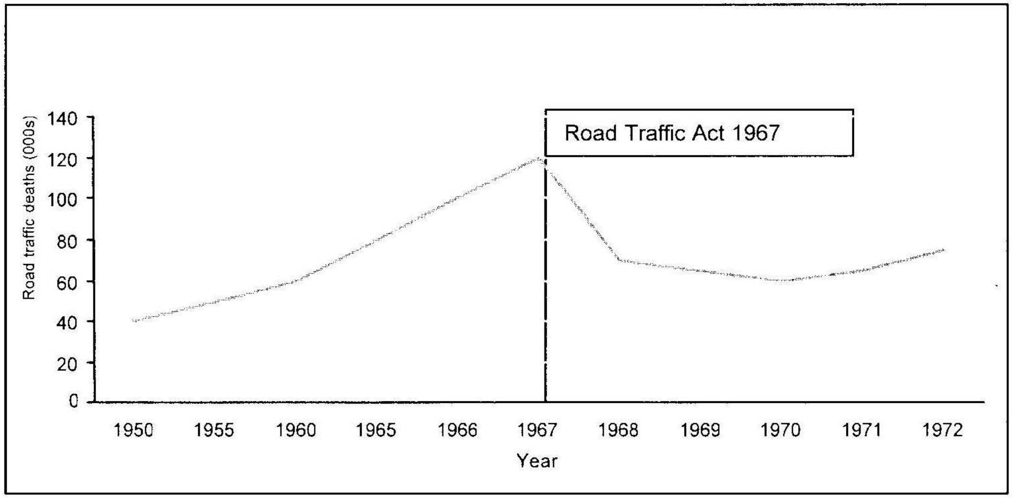
Panel study, QE logic with passive observation!

Is "parallel trends" assumption justifiable?

Time Series Studies

- "*Trend Analysis*": Look at one case over a long period of time, correlate changes in Y with changes in X that take place over the long time horizon. Examples: Does presidential popularity depend on economic performance? Did US military expenditures during the Cold War depend on USSR expenditures?
- "*Intervention Analysis*": Look at one case over a long time period, with a definite "intervention" or change in X that takes place sometime in that period. Did change in speed limit in country/state lead to fewer traffic fatalities?
- "*Multiple Time Series*": Intervention analysis with a control group

Figure 5: Interrupted time series design: road traffic deaths UK (1950 to 1972)



- This is an "intervention" analysis. Did 1967 law change the level of road fatalities in the UK? Strengths: Looks at long term trends and sees how the intervention altered it
- Problems:
  - History: What else might have happened in 1967?
  - Selection-Maturation: For example, UK was building up public frustration with fatalities (Z), which led it to adopt the new law but which also would have led to the decline in fatalities anyway
- Add control units with "matched countries" for greater confidence

- Final observational design in common use in political science: designs that include data gathered on units at **multiple levels of analysis**
  - E.g., individuals ("micro") within countries ("macro"), such as cross-national survey data; Political parties within regions within countries
  - Can extend this to longitudinal cases also so: individuals within countries within years (i.e., different individuals in the same countries interviewed at multiple occasions)
  - This is the format of some of the data you can analyze for your semester quantitative papers – from the World Values Data set
  - We call this kind of data "multilevel" or "hierarchical" data because units can be described as "nested" within higher-level units
- See Dalton *et al.* (2009) on syllabus as example
- Internal validity issues here as in other observational designs, but many advantages nevertheless

- Advantages of Multilevel Observational Designs
  - Strong **external validity**, given large numbers of observations from multiple sites
  - Can examine whether micro explanations differ from macro explanations, e.g. income *increases* likelihood of Republican voting at the micro-level but *decreases* it at macro (state/county)-levels
  - Can examine how **contextual** factors – at higher levels – influence outcomes, and whether they influence the way that individual-level processes work in different places
  - E.g. Cross-nationally, individuals may vote according to their "pocketbooks", but they are *more likely* to do so when there is a majority-party government than a coalition government (because people can more easily attribute responsibility or blame for their financial condition to the incumbent party in power
  - Macro-micro linkages and interactions are **hugely** important in contemporary political science!!!