Research Design in Political Science
POLS4011/POLS8058

Richard Frank
28 April 2026

WEEK 8: CAUSAL INFERENCE, PART 2 (SOME SOLUTIONS)

PART 1: OVERVIEW

In Week 7 we worked through why causal inference is hard. The fundamental problem of causal inference, as introduced through Morgan and Winship (2015), is that we can never observe the same unit in both its treated and untreated state at the same time, so every causal claim rests on a counterfactual that must be constructed or argued for. Mahoney (2008) complicates that picture by showing that case-level and population-level approaches to causation are connected through the logic of INUS and SUIN causes (wee!), and Cho, Dreher, and Neumayer (2013) provide a concrete example of how difficult it is to make credible causal claims from cross-sectional observational data alone. This week we move from focusing on the problem to several possible solutions.

The central message this week is that the various research design strategies in political science (e.g., experiments, natural experiments, difference-in-differences, instrumental variables, regression discontinuity, and process tracing) are best understood as different strategies for constructing credible counterfactuals. No single design is "best" for everything. Each approach involves making assumptions that are more or less plausible in context, and each involves trade-offs between internal validity, external validity, and feasibility. What matters is matching the design to the research question and being transparent about what the chosen design can and cannot support.

A pedagogical note. The emphasis this week is on design logic, not on statistical mechanics. You do not need to be able to derive the instrumental variable (IV) estimator from scratch to appreciate what an IV is doing. You do need to understand what each design assumes, what counterfactual it constructs, and when each is appropriate. If you leave this week able to read a paper and say, in plain language, "this design works because X, and its main vulnerability is Y," you will have gotten what you should from this week.

This week also continues the integration of quant and qual traditions that Mahoney (2008) began last week. Process tracing is sometimes treated as entirely separate from (and sometimes antithetical to) the quantitative causal inference toolkit, but Mahoney (2008) and others argue that the underlying logic is similar. Both are strategies for ruling out alternative explanations for an observed outcome, and both do so by making (ideally clear, specific, and testable) predictions about what should be true in the world if the causal story is correct. Bennett and Checkel (2015) formalise this by articulating best practices for process tracing, including what they call hoop tests and smoking gun tests, which function as evidentiary standards that parallel the identifying assumptions of quant designs. Dunning (2008) makes a related point from the other direction: natural experiments sit on a continuum between true randomisation and pure observation, and their credibility depends on how plausibly "as-if random" the assignment of the treatment is.

Plan for today

- Overview: from the problem to the solutions
- Readings: experimental and quasi-experimental designs, instrumental variables, and process tracing
- Group activity: a research design tournament
- Wrapping up the causal inference fortnight and looking ahead

Key themes for this week

- Research designs are strategies for constructing credible counterfactuals; each is a different solution to the fundamental problem of causal inference.
- Randomised experiments as the benchmark: why randomisation solves the selection problem, and why experiments are not always available or appropriate.
- Natural experiments and the "as-if random" assumption; Dunning's (2008) continuum of plausibility.
- Instrumental variables: the logic of exogenous variation, and the two core assumptions of relevance and exclusion.
- Difference-in-differences and regression discontinuity as alternative identification strategies.
- Process tracing and congruence testing: the qualitative tools for within-case causal inference, and their parallels to the quantitative toolkit.
- The applied example: Miguel, Satyanath, and Sergenti (2004) as a model application of IV logic, and a case study of where IV assumptions strain.

The differentiated expectations continue. Honours students should be able to identify what each major design assumes and how each constructs a counterfactual, and to recognise the main threats to each. MA/PhD students should be able to evaluate design choices in published research, articulate why a given identifying assumption is or is not credible in a specific context, and defend their own design against the most plausible alternatives.

PART 2: READINGS

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2): 282–293.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*, Ch. 1–2. Princeton, NJ: Princeton University Press.

Bennett, Andrew, and Jeffrey T. Checkel. 2015. "Process Tracing: From Philosophical Roots to Best Practices." In Andrew Bennett and Jeffrey T. Checkel, eds., *Process Tracing: From Metaphor to Analytic Tool*. New York: Cambridge University Press.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112(4): 725–753.

Dunning (2008)

Dunning (2008) is the conceptual anchor for our quantitative design discussion this week. His goal is to take stock of the growing use of natural experiments in political science and offer a framework for evaluating them. This is a much shorter treatment of the subject than his 2012 book, although I recommend the book if you are interested in exploring natural experiments in your own research. The key move in the article is to locate natural experiments on a continuum, rather than treating them as a single, homogeneous category of research design. At one end of the continuum sit true randomised experiments, in which the researcher controls treatment assignment and randomises it. At the other end sit purely observational studies, in which treatment assignment may be correlated with all sorts of confounders. Natural experiments sit between these poles, and the claim that a given study is a "natural experiment" depends on how plausibly the treatment assignment is argued to be as-if random.

This continuum framing does a lot of work. It pushes back against the tendency, which was more common in the mid-2000s when Dunning was writing, to treat the label "natural experiment" as a credential that absolves the researcher of further justification. Dunning (2008) argues that the phrase describes a claim about the data-generating process, not an intrinsic property of a dataset. A study exploiting an electoral redistricting is a natural experiment only to the extent that the boundaries between treated and untreated units are genuinely unrelated to the outcome we want to explain. When the assignment is plausibly as-if random, natural experiments can approach the credibility of true experiments. When it is not, they are observational studies in disguise.

Dunning (2008) distinguishes three main types of natural experiment. First, standard natural experiments, where some naturally occurring process (e.g., a policy cut-off, a geographic boundary, or a lottery) generates variation in treatment that is plausibly independent of potential outcomes. Hyde (2007) is one of his examples. Second, regression discontinuity designs, where treatment is assigned based on whether a unit falls above or below a threshold on some running variable, and the researcher exploits the fact that units just above and just below the threshold should be similar in all respects except treatment. Third, instrumental variables, where the researcher finds a variable that affects the treatment but has no direct effect on the outcome, and uses it to isolate exogenous variation in the treatment. Miguel, Satyanath, and Sergenti (2004), our applied example this week, is Dunning's (2008) IV example.

The strengths of natural experiments follow from the as-if random assumption. When it holds, the simple comparison of treated and untreated units provides a credible estimate of the causal effect, without the heavy reliance on regression controls that plague observational research. The identifying assumptions, while demanding, are usually easier to state and to evaluate than the long list of "no unobserved confounders" assumptions that conventional regression requires.

The limitations are equally important and are where I would encourage you to spend most of your effort when you read the article. First, natural experiments often estimate local average treatment effects (LATE), not average treatment effects for the population. The estimate reflects the causal effect for the specific subset of units affected by the natural experiment, which may not generalise. Second, the as-if random assumption is frequently more demanding than authors acknowledge. A geographic discontinuity, for example, only identifies a causal effect if units on either side of the border are similar in all respects except the treatment, which is rarely strictly true. Third, natural experiments typically sacrifice external validity for internal

validity. A credible natural experiment in one country at one time (e.g. Armenia in 2003) may tell us a great deal about that context and very little about anywhere else. Fourth, the very scarcity of natural experiments means the questions we can ask are constrained by the variation that history happens to have produced, not by substantive importance.

The upshot is that natural experiments are a powerful addition to the political science toolkit, but they are not a substitute for careful thinking about research design. Dunning's (2008) continuum framing is a useful corrective to the tendency (then and now) to treat "natural experiment" as a shield rather than an argument. When reading natural experimental studies, your job is to interrogate the as-if random assumption as hard as the authors do, and often harder.

Reading questions

Honours

1. Dunning (2008) argues that natural experiments lie on a continuum of plausibility. In your own words, explain what determines where on that continuum a given study sits. Use an example from the Week 5 or Week 6 readings.
2. What is the as-if random assumption, and why is it the central assumption in any natural experiment? What kinds of evidence would convince you that it holds in a particular study?

MA/PhD

1. Dunning (2008) identifies three types of natural experiment (standard, regression discontinuity, and instrumental variables). For each, state precisely what the identifying assumption is and what kind of evidence you would need to assess it. Which of the three do you regard as the most demanding and why?
2. Dunning (2008) argues that natural experiments often estimate local average treatment effects rather than average treatment effects. What are the implications of this distinction for how findings from natural experiments should be interpreted and generalised? Think of an example where the distinction would be substantively important.

Angrist and Pischke (2009), Ch. 1 and Ch. 2

Angrist and Pischke (2009) give you the econometric foundation for quantitative causal inference. These two chapters are the most accessible statement that I could find of the design-based causal inference paradigm that has come to dominate applied empirical work in economics and political science. If Morgan and Winship (2015) gave you the formal potential outcomes notation last week, Angrist and Pischke (2009) give you the applied sensibility: what kinds of research questions are amenable to credible causal inference, and what kinds of designs answer them.

Chapter 1 argues that every good empirical research question has, at its heart, a well-defined thought experiment. Before worrying about data or methods, the researcher should be able to specify: what is the treatment, what is the counterfactual, and in an ideal world, what experiment would settle the question? This is what they call the FAQ approach (Frequently Asked Questions): what is the causal relationship of interest, what experiment would identify the causal effect if cost were no object, what is the identification strategy, and what is the mode of statistical inference? These four questions are a useful checklist. Many papers in political

science that struggle to establish causal claims do so because they have not answered them clearly before collecting data. The chapter is also a good antidote to the temptation to think that more sophisticated statistics can substitute for a clearer research design. They cannot.

Chapter 2 introduces the experimental ideal and uses it to motivate the whole design-based approach. The core argument is that the randomised experiment is the benchmark against which all other designs should be evaluated, not because experiments are always feasible (they often are not), but because the logic of randomisation provides the cleanest possible solution to the selection problem. When treatment is randomly assigned, treated and untreated units are, in expectation, identical on all characteristics except treatment. The simple difference in means is then an unbiased estimate of the average treatment effect. No regression controls, no heroic assumptions about unobservables, no econometric gymnastics. Randomisation does the work.

From this benchmark, the chapter derives a lesson that runs through the rest of the book (which is also worth reading if you are interested and have the time). Observational research is credible to the extent that it approximates the conditions of an experiment. A natural experiment is credible when the assignment is as-if random. A regression discontinuity design is credible when units near the threshold are similar to those receiving one treatment or the other. An instrumental variable is credible when it isolates exogenous variation in the treatment. Each of these designs is a strategy for recovering something close to the experimental ideal in a setting where true randomisation is not available.

The chapter also introduces the key technical problem that motivates the rest of the book: selection bias. If treatment is not randomly assigned, the difference in means between treated and untreated units reflects not only the causal effect of the treatment but also the systematic differences between who is treated and who is not. Controlling for observable confounders helps, but only if the untreated units are a good counterfactual for the treated units on all relevant dimensions (observable and unobservable). This is the conditional independence assumption, and it is demanding. Most of this book is about research designs that relax or replace it with more credible assumptions.

Two things are worth noting about how Angrist and Pischke (2009) write. First, their tone is applied and often a bit cheeky, but the underlying arguments are careful and worth taking seriously. Second, they are unapologetic partisans of the design-based approach, and they are sometimes dismissive of alternatives (structural modelling, qualitative work, theory-driven approaches). You do not have to accept their full worldview to benefit from the chapters. The discipline they impose (ask a clear causal question, specify the ideal experiment, evaluate your design against it) is valuable whether or not you end up doing design-based empirical work.

Reading questions

Honours

3. Angrist and Pischke (2009) argue that every empirical research question should be accompanied by a clear statement of the ideal experiment. For your own research project, what is the ideal experiment? If it is not feasible, what is the main obstacle?
4. Why is randomisation such a powerful tool for causal inference? Explain in your own words how randomisation solves the selection problem, and why observational research generally does not.

MA/PhD

3. Angrist and Pischke (2009) place randomised experiments at the centre of their causal inference framework and treat other designs as approximations to the experimental ideal. Is this framing always appropriate? Are there research questions for which the experimental ideal is not the right benchmark, or where the logic of experimentation breaks down even in principle?
4. The conditional independence assumption (that treatment is independent of potential outcomes conditional on observables) underlies regression-based causal inference. What would it take to convince you that this assumption holds in a specific applied setting? Can you think of a political science example in which you find the assumption credible, and one in which you do not?

Bennett and Checkel (2015)

Bennett and Checkel (2015) give us the qualitative side. Their chapter is the introduction to an edited volume on process tracing and, along with Mahoney (2008) from last week, it is the most important statement I could find on how qualitative research can contribute to causal inference. Read alongside Dunning (2008) and Angrist and Pischke (2009), it makes visible a point that quantitative methodologists often miss: process tracing is not a soft substitute for causal identification. It is a disciplined strategy for testing causal claims within cases, with its own evidentiary standards and its own rules for what counts as evidence.

The chapter begins by defining process tracing as the analysis of evidence on processes, sequences, and conjunctures of events within a case for the purpose of either developing or testing hypotheses about causal mechanisms. Two features of the definition deserve attention. First, process tracing is explicitly about causal mechanisms, not just causal effects. Where quantitative designs typically estimate "the effect of X on Y," process tracing asks *how* X produces Y, and what intermediate steps should be observable if the causal story is correct. Second, process tracing is about cases, not populations. It is an inferential strategy appropriate for single cases or small numbers of cases, where the logic of statistical comparison does not apply.

The conceptual core of the chapter is the typology of process tracing tests, which Bennett and Checkel (2015) borrow from Van Evera (1997). There are four types of test, each defined by two properties: whether passing the test is necessary for the hypothesis to be true, and whether passing it is sufficient to confirm the hypothesis. A straw-in-the-wind test is neither necessary nor sufficient: passing or failing it shifts the credibility of the hypothesis modestly. A hoop test is necessary but not sufficient: passing it is required for the hypothesis to survive, but passing does not confirm it. Failing a hoop test, however, is fatal. A smoking-gun test is sufficient but not necessary: passing it strongly confirms the hypothesis, but failing it does not refute it. A doubly decisive test is both necessary and sufficient: passing it confirms the hypothesis and failing it refutes it. These tests are rare in practice, but the typology is useful as a way of thinking about the evidentiary weight of specific pieces of evidence.

The parallel with quantitative causal inference should be clear. A hoop test is analogous to a necessary condition for a design to be credible: if the hoop test fails, the causal claim cannot stand. A smoking-gun test is analogous to direct evidence of a mechanism: if it holds, the causal claim is strongly supported. The difference is that quantitative designs rely on statistical assumptions about the data-generating process (e.g., as-if random assignment, exclusion restriction, and conditional independence), while process tracing relies on substantive

predictions about what should be observable in the case if the causal story is correct. Both strategies are about ruling out alternative explanations. They differ in the kinds of evidence they focus on.

The heart of the chapter is the list of ten best practices for process tracing. I will not rehash all ten here, but several deserve stressing. First, cast the net widely for alternative explanations. This is the qualitative analogue of the quantitative researcher's search for confounders, and it is the single most important discipline in process tracing. A process trace that considers only the author's preferred explanation is not process tracing; it is confirmation of priors. Second, be equally tough on alternative explanations as on your own. This is harder than it sounds, and Bennett and Checkel (2015) are frank that most process tracing in print does not meet the standard. Third, consider the potential biases of sources. Documents, interviews, and secondary sources all have biases, and a process trace that takes them at face value is weaker for it. Fourth, take into account whether the case is most or least likely for alternative explanations. A case that is a least-likely case for your explanation but where your explanation nonetheless holds provides stronger evidence than a case where your explanation was always going to look good. These practices echo the case selection logic we covered in Weeks 5 and 6.

The chapter also makes an important philosophical point about the relationship between process tracing and Bayesianism. Process tracing is, at heart, a Bayesian procedure. We begin with prior beliefs about the relative credibility of competing hypotheses, observe evidence, and update our beliefs accordingly. The strength of the updating depends on how likely the evidence would be under each hypothesis. A smoking-gun test is powerful because the evidence would be very unlikely under alternative hypotheses. A hoop test is a strong negative test because failing it would be highly unlikely under the hypothesis of interest. You do not need to do formal Bayesian math to do good process tracing, but thinking Bayesianly (if that is a word) about evidence is helpful.

For this class, Bennett and Checkel (2015) is the reading that could reshape how you think about qualitative research. If you are planning a qualitative or mixed-methods project, the best practices list is genuinely useful as a checklist when you are writing. If you are planning a quantitative project, the chapter is still valuable because it shows you what it looks like to evaluate causal mechanisms in specific cases, which is something you may well want to do alongside a regression, an experiment, or an IV estimation. The move from Mahoney (2008) last week to Bennett and Checkel (2015) this week is from the conceptual architecture of case-level causation to the practical toolkit for doing case-level causal inference well.

Reading questions

Honours

5. Explain in your own words the difference between a hoop test and a smoking-gun test. Can you think of an example, from a topic you are familiar with, where a specific piece of evidence would function as one or the other?
6. Bennett and Checkel (2015) argue that process tracing requires casting the net widely for alternative explanations. For a causal claim you have read or made in your own work, list three alternative explanations and briefly say what evidence would help distinguish between them.

MA/PhD

5. Bennett and Checkel (2015) present process tracing as a Bayesian procedure. Explain what this means and evaluate the claim. Does thinking of process tracing as Bayesian updating help or hurt its credibility as a causal inference strategy?
6. How do the Bennett and Checkel (2015) best practices for process tracing relate to the Dunning (2008) framework for natural experiments? Are they alternative strategies for the same inferential goal, complementary strategies for different goals, or are they doing fundamentally different things?

Miguel, Satyanath, and Sergenti (2004)

Miguel, Satyanath, and Sergenti (2004) is our applied example this week, and it is one of the most widely cited applications of instrumental variables in political science. The research question is whether negative economic shocks cause civil conflict in Sub-Saharan Africa. The challenge is that economic performance and conflict are jointly determined (and thus endogeneous): conflict destroys economic activity, and poor economic performance may encourage conflict, so a simple regression of conflict on growth cannot distinguish cause from effect. The authors' strategy is to instrument for growth using variation in rainfall, which in their sample strongly predicts growth (because most African economies are heavily rain-fed agricultural) and which, they argue, affects conflict only through its effect on the economy. If both assumptions hold, the IV estimate recovers the causal effect of economic shocks on conflict.

The article is a clean illustration of the IV logic. Instrumental variables rest on two core assumptions. The first is relevance: the instrument must affect the treatment. Miguel, Satyanath, and Sergenti (2004) devote considerable space to the first stage regression of growth on rainfall, showing a strong, statistically significant relationship. Rainfall variation explains a meaningful share of the variation in growth in their sample. Relevance is the easy assumption to evaluate because it is testable directly from the data.

The second assumption, the exclusion restriction, is the hard one. The instrument must affect the outcome only through the treatment, not through any other channel. The authors argue that rainfall affects conflict only through growth: the only reason rainfall should matter for civil conflict is because it changes economic performance, and thus changes the opportunity cost of joining a rebellion. This assumption is not directly testable. The authors defend it with a series of arguments: rainfall is plausibly unrelated to institutional quality, to ethnic composition, to prior conflict, and to other structural determinants of civil war. They also conduct a range of robustness checks. But the exclusion restriction is an assumption, not a demonstrated fact, and the credibility of the IV estimate depends on it.

This is exactly where the design becomes interesting for our purposes. You should read the article thinking about Dunning's (2008) continuum. Miguel, Satyanath, and Sergenti (2004) are making a case that their IV satisfies the exclusion restriction, and they are locating their study close to the as-if random end of the continuum. Is their argument convincing? There are plausible alternative channels through which rainfall might affect conflict: directly, through competition over water or land; through migration patterns driven by drought; through food security and nutrition in ways not fully captured by GDP growth; or through the logistical feasibility of military operations. Each of these is an alternative channel that, if operative, would violate the exclusion restriction and bias the IV estimate. The authors address some of these concerns and not others. A subsequent literature (which you will encounter if you work

in this area) has been critical of the rainfall-as-instrument strategy on exactly these grounds. Schultz and Mankin (2019) and others have pushed back on the empirical robustness and the exclusion restriction.

The main substantive finding is large: a five percentage-point drop in economic growth is associated with roughly a 50 percent increase in the likelihood of civil conflict in the following year. Whether you find this finding credible depends on whether you find the exclusion restriction credible. If you do, the IV estimate is an unbiased causal effect. If you do not, the estimate is biased by the alternative channels through which rainfall affects conflict, in a direction that depends on the specifics. This is why IV studies need to be read with care: the coefficient reported is only as credible as the instrument, and evaluating the instrument is usually a matter of substantive knowledge about the context rather than statistical tests.

I have assigned this article because it is an excellent example of what a careful, transparent IV study looks like. The authors state their assumptions clearly, conduct the robustness checks the assumptions suggest, and acknowledge the limitations. They are not overclaiming. But the article also illustrates a point that you should carry with you: even a careful, transparent IV study stands or falls on the exclusion restriction, which cannot be tested directly. The most important skill when reading IV papers is being able to ask, specifically, what other channels could connect the instrument to the outcome, and to evaluate how the authors have (or have not) addressed them. The answer is rarely "none." The question is whether the other channels are small enough or well-enough accounted for that the estimate remains informative.

Reading questions

Honours

7. State the two core assumptions of any instrumental variables design in your own words. For Miguel, Satyanath, and Sergenti (2004), which of the two is more credible and why?
8. Miguel, Satyanath, and Sergenti (2004) claim that rainfall affects civil conflict only through its effect on economic growth. List two plausible alternative channels through which rainfall might affect conflict, and assess whether they would bias the IV estimate upward or downward.

MA/PhD

7. The exclusion restriction cannot be tested directly. What evidence do the authors marshal to defend it, and how convincing is that evidence? What additional evidence (quantitative or qualitative) would strengthen the case?
8. Miguel, Satyanath, and Sergenti (2004) estimate a local average treatment effect: the effect of growth on conflict for those countries and years in which growth is driven by rainfall variation. What does this mean for how the finding should be generalised? In which contexts would you expect the effect to be larger or smaller than what the IV estimate recovers?

Cross-reading question

1. Dunning (2008), Angrist and Pischke (2009), and Bennett and Checkel (2015) each propose a set of standards for credible causal inference. Are they proposing alternative standards for the same inferential goal, or different standards for different goals? How would you decide which set of standards applies to a given research question?

PART 3: GROUP ACTIVITY

The Research Design Tournament

Last week's group activity put you in the role of a critic, diagnosing causal claims in published research. This week flips the task. You will now be the designer. Each group will be assigned a research design strategy from this week's readings and asked to build, from scratch, a credible causal inference study using only that design. At the end, each group will pitch its design to the class, and together we will assess which designs are most credible for the question in front of us and why.

The pedagogical point here is that you do not fully understand a research design until you can use it. Reading about IV is not the same as designing an IV study. Reading about process tracing is not the same as specifying what evidence you would look for and what tests it would constitute. Working through the design from the inside will make the logic stick.

The setup

I will divide the class into six groups. Each group will be assigned one of the following research design strategies. The six strategies correspond to the main tools covered in this week's readings:

- Group A: Randomised field experiment
- Group B: Natural experiment (standard, not RDD or IV)
- Group C: Regression discontinuity design (RDD)
- Group D: Difference-in-differences (DiD)
- Group E: Instrumental variables (IV)
- Group F: Process tracing

All groups will work on the same substantive question. You have a choice of three. Pick one within your group, but all groups in the room should aim to work on the same question so the designs are directly comparable. I will rotate questions across sections if needed.

The research questions (pick one)

1. Does foreign aid reduce the incidence or duration of civil conflict in recipient states?

2. Does international electoral observation reduce electoral fraud?

3. Does legalising prostitution affect the inflow of human trafficking?

You will notice that each question is drawn from readings you have already encountered in this course: Fortna (2004), Hyde (2007), and Cho, Dreher, and Neumayer (2013). That is on purpose. You already know something about each substantive topic, so you can hopefully concentrate on the design logic rather than on getting up to speed on a new literature.

Your task

Working in your groups, design a study of your chosen research question using only your assigned research design strategy. Work through the following steps. You have about 25 minutes, then each group will present a five-minute pitch to the class.

Step 1: Specify the design
What, concretely, is your study? What is the treatment, who are the units, what is the outcome, and how is the treatment assigned (or, for process tracing, what is the case and what mechanism are you tracing)? Give a specific enough description that another group could, in principle, execute your design. For the quantitative designs, specify what data you would need and where it might come from. For process tracing, specify the case, the candidate causal mechanism, and the kinds of evidence you would gather.

Step 2: Articulate the counterfactual
What counterfactual is your design constructing, and how? For the experimental and quasi-experimental designs, what comparison group stands in for the unobserved counterfactual? Why is the comparison credible? For process tracing, what alternative explanations are you ruling out, and what evidence would allow you to rule them out?

Step 3: State the identifying assumption
Every design rests on at least one key assumption that cannot be tested directly from the data. State it clearly. For an experiment, it is the integrity of randomisation and compliance. For a natural experiment, it is as-if randomness. For RDD, it is the continuity of potential outcomes at the threshold. For DiD, it is parallel trends. For IV, it is relevance and the exclusion restriction. For process tracing, it is that the evidence you would gather discriminates between your preferred explanation and the leading alternatives. State the assumption in a form that could, in principle, be contested.

Step 4: Identify the main threat
What is the single most plausible reason your design might fail? This is the threat you expect an informed critic (i.e., the other groups in the room) to raise. Be honest. A design you cannot think of a serious threat to is probably a design you have not thought about hard enough.

Step 5: Pitch
Make the case for why your design is the best available strategy for answering the research question, given realistic constraints on data, time, and feasibility. What does your design deliver that the others cannot? Where are its limitations?

The pitch
Each group will have about five minutes to present. Structure the pitch as follows. First, state the design in one sentence. Second, describe the counterfactual and the identifying assumption. Third, identify the main threat and what, if anything, can be done about it. Fourth, make the case for why your design is appropriate for the question. Expect the rest of the class to press you on the identifying assumption and on the main threat.

After each group has presented, we will discuss as a class. Which designs are most credible for the question? Which are least feasible? Where would a combined design (say, an IV supplemented with process tracing) be stronger than either alone? The tournament framing is a device, not a real competition. The point is to see six different designs applied to the same question and to think carefully about what each contributes.

A note on how to approach this
You may find that your assigned design is awkward for your research question. That is useful information. The designs in your toolkit are not universally applicable. Some questions are

easier to answer with one design than another, and some questions may simply not be answerable with any single design. Being able to articulate why is a core skill.

PART 4: WRAPPING UP AND LOOKING AHEAD

Let me try and pull the causal inference fortnight together. Last week we established the problem: we cannot observe the same unit in both its treated and untreated states, so every causal claim rests on a counterfactual that must be constructed or argued for. Mahoney (2008) and Morgan and Winship (2015) gave us two complementary frameworks, one case-oriented and one population-oriented, for thinking about what that counterfactual means.

This week we moved from the problem to the solutions. Randomised experiments are the benchmark because randomisation, when it works, solves the selection problem by construction. Natural experiments, regression discontinuity, difference-in-differences, and instrumental variables are each strategies for approximating the experimental ideal in settings where true randomisation is not available. Each rests on assumptions that must be stated clearly and defended substantively, and each is more or less credible depending on how plausible those assumptions are in context. Dunning's (2008) continuum of plausibility is a useful way of thinking about all of them. On the qualitative side, process tracing provides a disciplined strategy for causal inference within cases. Bennett and Checkel's (2015) best practices and typology of tests show that process tracing is not a soft substitute for causal identification but a separate, rigorous strategy with its own evidentiary standards.

Three broader takeaways from the fortnight. First, causal inference is an exercise in ruling out alternative explanations. Every design does this differently, but the goal is the same. The quantitative designs rule out alternatives through assumptions about the data-generating process (as-if random assignment, exclusion restriction, parallel trends). Process tracing rules out alternatives through substantive predictions about what should be observable if each candidate explanation is true. Both are disciplined, both require hard thinking about what could go wrong, and both are strengthened by being explicit about the alternatives they are trying to rule out. Second, no single design is best. The right design depends on the question, the available data, and the context. A randomised experiment is powerful when feasible and appropriate; it is useless when the treatment cannot be randomised for ethical or practical reasons. An IV is powerful when a credible instrument exists; it is meaningless when the exclusion restriction is implausible. Process tracing is powerful when the case is well documented and the mechanism is clearly articulated; it is unhelpful when the causal story is vague or the evidence is thin. Third, mixing designs is often stronger than relying on any single one. An IV supplemented with process tracing of one or two cases may be more convincing than either alone, because the qualitative evidence can help evaluate the exclusion restriction in a way that statistics cannot.

The payoff of this fortnight is that you should now be able to read a piece of political science research and articulate, with some precision, what causal claim is being made, what counterfactual the design is constructing, what the identifying assumption is, and where the design is most vulnerable. These are not abstract methodological concerns. They are the core of what it means to do credible political science research, and they will show up in every empirical paper you read and every defence of your own research you give.