Research Design in Political Science
POLS4011/POLS8058

Richard Frank
12 May 2026

WEEK 10: WHEN THINGS FALL APART, PART 2 (ROBUSTNESS & OPEN SCIENCE)

PART 1: OVERVIEW

Last week we developed a diagnostic vocabulary for threats to inference including a four-fold validity typology, omitted variable bias, the limits of pattern-finding, etc. This week we shift from *diagnosis* to *response*. The two questions are different. Diagnosis asks, "what could go wrong with this design?"; response asks, "given what could go wrong, what do you actually do about it, in a way that is credible to a sceptical reader?" Both questions must be answered in any honest empirical paper, and the move from one to the other is not automatic. A long list of threats with no responses is analytical paralysis; a long list of responses with no diagnosis is research theatre.

The substantive content this week falls into three blocks. The first is robustness as a design and analytical practice including sensitivity analyses, placebo tests, bounding exercises, and triangulation across designs. The Plümper and Traunmüller (2020) paper is the centrepiece here. The second block is the open-science apparatus that the social and behavioural sciences have built up over the past fifteen years, including pre-registration, pre-analysis plans, registered reports, the disclosure standards being adopted by journals, and the data-sharing practices that have moved from optional to expected in many subfields. Miguel et al. (2014) is the policy-defining statement, and Tyner et al. (2026) is the most recent large-scale evidence on what these efforts have, and have not, achieved. The third block is measurement of the conditions that make the open-science apparatus possible at all: Hollyer et al. (2014) on transparency, which I read as a methodological work on construct validity as much as a substantive piece on government behaviour. Bazzi and Blattman (2014) is this week's applied paper as a companion to Week 8's Miguel et al.'s (2004) instrumental variable paper. There is the same broad question, similar setting, but the conclusions shift once additional controls, samples, and outcome definitions are introduced. It is the cleanest demonstration, I think, this semester of why robustness checks matter.

This is the last substantive week before presentations begin. We have spent ten (!) weeks building up the vocabulary, the design tools, and the diagnostic mentality that empirical work in political science requires. The next two weeks ask you to apply that apparatus to your own project, in front of an audience that has been through the same material. Today is the moment to take stock. I want you to leave this week with a working answer to three questions: what do the readings this week tell me about how to defend my own design? What does this semester tell me about how to make a credible empirical argument? What am I going to do, in my presentation and proposal, with the threats I have not yet solved?

Plan for today

- Overview: from diagnosing threats to responding to them.
- Readings: the limits of sensitivity analysis, the measurement of transparency, the open-science programme, the most recent replicability evidence, and a worked replication-and-extension.
- Class discussion: drawing the readings into conversation, drawing the semester together, and looking forward to the presentations.
- Wrapping up the semester.

Key themes for this week

- Robustness is a property of your theoretical argument, not a regression button you push in R. The same data, with a different definition of robustness, can yield a different verdict.
- Pre-registration draws a useful line between confirmatory and exploratory analysis, but the line is harder to maintain in practice than in principle, and the cost of pre-registration varies by subfield.
- Transparency is layered: of the data, of the code, of the analytical choices, and of the reasoning behind them. Each layer addresses a distinct threat to credibility.
- Replicability across the social and behavioural sciences is heterogeneous. Reading any new evidence carefully matters more than remembering the headline numbers.
- Honest reporting of limitations is the practice that ties this week to last week. The diagnostic vocabulary from Week 9 is what makes the limitations section of your papers credible; the response toolkit from this week is what makes it useful.

The differentiated expectations carry through to the end. Honours students should be able to identify the most credible robustness checks and transparency practices for their own design, and to articulate how they would pre-register their analysis if they chose to. MA and PhD students should be able to evaluate the transparency, robustness, and replicability claims in published research, and to defend their own response strategy against the most plausible alternatives.

PART 2: READINGS

Plümper, Thomas, and Richard Traunmüller. 2020. "The Sensitivity of Sensitivity Analysis." *Political Science Research and Methods* 8(1): 149-159.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2014. "Measuring Transparency." *Political Analysis* 22: 413-434.

Miguel, E., et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343(6166): 30-31.

Tyner, Andrew H., et al. 2026. "Investigating the Replicability of the Social and Behavioural Sciences." *Nature* 652: 143-150.

Bazzi, Samuel, and Christopher Blattman. 2014. "Economic Shocks and Conflict: Evidence from Commodity Prices." *American Economic Journal: Macroeconomics* 6(4): 1-38.

Plümper and Traunmüller (2020), "The Sensitivity of Sensitivity Analysis"

Plümper and Traunmüller (2020) is short, quietly devastating (to me at least), and a useful corrective to the way "sensitivity analysis" is invoked in applied work. The standard practice is to take the main regression, change the model in some way (drop a control, add a control, change the sample, change the standard errors, switch the estimator, etc.), and report whether the coefficient of interest survives. If it does, the finding is "robust." If it does not, the analyst chooses which of the specifications to feature in the paper, and the others get dumped in an online appendix. The implicit logic is that survival across many specifications is evidence of a real effect. Plümper and Traunmüller (2020) show, with Monte Carlo evidence, that this logic does not hold up.

Their argument has several parts that are worth considering. First, what counts as "robust" is not a property of the world; it is a definition. The literature uses several definitions (extreme bounds analysis, Bayesian model averaging, the Sala-i-Martin (1997) model averaging rule, simple sign-and-significance survival across specifications), and these definitions lead to different substantive conclusions, in finite samples, on the same data. A coefficient classified as robust by one definition can be classified as fragile by another. The choice of definition therefore makes a substantive difference, and it is rarely justified in published research. Even setting the choice of definition aside, the validity of any sensitivity test depends on properties of the data-generating process that the researcher cannot directly observe, particularly with the correlation structure between determinants and confounders. When determinants and confounders are nearly uncorrelated, sensitivity tests perform reasonably well. As that correlation rises, which is the case in the messy observational settings where sensitivity tests are most often invoked, the tests deteriorate. Also, sensitivity tests can usefully detect false positives (a variable that survives extreme bounds is almost certainly in the model), but they are unreliable for detecting fragility. Lack of robustness across specifications is not, on its own, evidence that the underlying claim is wrong; it may instead be evidence that the scholar is changing the bias structure across specifications, in the sense that Clarke (2005) warned us about last week.

The constructive implication of this article is that "I ran a battery of robustness checks and the result survived" is not, on its own, a strong claim. The same battery, on the same data, with a different definition of robustness, may produce a different conclusion. The solution is not to abandon robustness checks completely but to be specific about what kind of robustness is being claimed, why that kind matters for the inference at issue, and what the test has the power to detect. This is consistent with the broader theme of this week: design-level transparency about what was done, and why, beats post hoc claims about result resilience.

Reading questions

Honours

1. In your own words, why does the choice of robustness test matter? Give an example, real or hypothetical, in which a coefficient might be classified as robust under one definition and fragile under another.
2. For your own research project, what would a credible robustness check look like? What property of the data-generating process would it have to address, and how would you justify the test you chose?

MA/PhD

1. Plümper and Traunmüller (2020) suggest that lack of robustness is not, on its own, evidence against a finding. Is that a defensible position? When should we treat fragility as informative?
2. How should journal editors and reviewers respond to the paper's argument? Should the standard for "robust" findings be raised, lowered, or restructured?

Hollyer, Rosendorff, and Vreeland (2014), "Measuring Transparency"

Hollyer, Rosendorff, and Vreeland (2014) is included this week for two reasons. The substantive contribution is the construction of the HRV index of government transparency, derived from an item response model that treats transparency as a latent predictor of countries' reporting of aggregate data to the World Bank's World Development Indicators. The methodological contribution is a model of how to build a measure of a latent construct in a way that is honest about its assumptions, estimable on real data, and useful for downstream causal inference. This connects directly to the conceptualisation and measurement readings from Weeks 3 and 4, and to the construct validity branch of the Shadish et al.c(2002) typology that we worked through last week.

The substantive payoff also matters. Existing measures of transparency, particularly those derived from Freedom House or from press freedom indices, are bundles. They mix outcomes (a free press, an informed public) with inputs (the disclosures governments choose to make), and they are coded by analysts whose judgments are difficult to validate against an external benchmark. The HRV measure narrows the construct deliberately. It captures one dimension only (the government's collection and dissemination of aggregate data), and it estimates that dimension from a behavioural trace (whether and when countries report each indicator) rather than from analyst judgment. The narrower construct is a feature, not a bug: the authors are explicit that transparency is multidimensional, and that the HRV measure speaks to one dimension that is theoretically interesting and that turns out to predict a range of political and economic outcomes once measured well.

The connection to the rest of today is that measurement is upstream of analysis in the same way that data generation is upstream of analysis (the Kalyvas [2004] point from last week). A research project that takes a noisy or bundled measure of its central concept is making a methodological commitment whether the analyst notices or not, and no downstream robustness check can repair a measurement choice that does not capture the construct. The Hollyer et al. (2014) approach is also a model of how to use a latent-variable framework to combine multiple observed indicators into a single measure with quantified uncertainty, which is a practice you will encounter often, and which several of you may want to use in your own work.

Reading questions

Honours

3. Hollyer et al. (2014) construct their measure by treating transparency as a latent predictor of an observable behaviour (data reporting). What are the strengths and limits of this approach? What would the index miss?

4. If you were measuring a latent concept in your own research, what observable behaviours would you use as indicators, and what assumptions would tying them to the latent concept require?

MA/PhD

3. Compare the HRV index with the Freedom House measure of press freedom. What are the implications of the construct difference for the kinds of causal claims that can be supported by each?
4. Item response models are increasingly common for political science measurement (V-Dem, ideal point estimation, and so on). What are the costs of widespread reliance on this family of models, and what alternatives exist when the assumptions break down?

Miguel et al. (2014), "Promoting Transparency in Social Science Research"

Miguel et al. (2014) is the short article that articulated the transparency agenda which much of the field has since adopted. The article is brief but ambitious. It calls for three sets of practices, each of which has since been operationalised, debated, and partially institutionalised. First, disclosure: systematic reporting standards (analogous to CONSORT in medicine) so that readers can identify what was done in a study. Second, registration and pre-analysis plans: ex ante specification of hypotheses and tests, with the explicit goal of disciplining the line between confirmatory and exploratory analysis. Third, open data and materials: making the underlying data, code, and replication files available so that the analytical claims can be checked and the analysis extended.

Each of these practices addresses a specific threat to the credibility of empirical findings, and each carries costs that are worth being clear about. Disclosure addresses the gap between what was done and what was reported, and it has costs in writing time and in the embarrassment of having to admit choices that look better unstated. Pre-registration addresses the file-drawer problem and the routine slide from exploration into confirmation, and it has costs in flexibility, particularly for projects whose data structure is not fully known in advance. Open data addresses the problem that findings that cannot be checked are findings that cannot be trusted, and it has costs that vary by subfield (almost none for survey experimentalists, substantial for ethnographers and for researchers working with sensitive populations). The honest presentation of the transparency agenda recognises these costs and asks how they are best managed, rather than insisting on a uniform standard that is wrong for some kinds of work.

For our purposes, Miguel et al. (2014) is best read as the policy framework against which to evaluate our own practices. The call is concrete: choose a journal that has implemented disclosure standards; pre-register analyses where the data structure permits; deposit code and data when ethics and access allow it. None of this is required for your degree here, and most of it is not required by the journals some of you are likely to publish in early in your career. The argument for adopting these practices anyway is that they raise the credibility of your own work, they reduce the time you will spend later defending it against the kinds of criticism we worked through last week, and they contribute to a scholarly culture in which the next generation of work has a higher floor.

Reading questions

Honours

5. Miguel et al. (2014) propose three transparency practices. Which of these would you adopt for your own project, and which would you find difficult to adopt? Why?
6. The article distinguishes between confirmatory and exploratory analysis. Why does this distinction matter, and how would you communicate it in a paper that does both (as most papers do)?

MA/PhD

5. Pre-registration is not free; it constrains the analyst in ways that may be costly when the data turn out to look different from what was anticipated. How should pre-registration be designed to discipline the analyst without making the analysis brittle?
6. Open data is straightforward in some subfields and impossible in others. Should the discipline have a single transparency standard, or should standards be calibrated to subfield? What are the costs of each option?

Tyner et al. (2026), "Investigating the Replicability of the Social and Behavioural Sciences"

Tyner et al. (2026) is the most current major piece of evidence on the state of replicability across the social and behavioural sciences I could find. The team attempted replications of 274 claims drawn from 164 quantitative papers, sampled from 54 journals across 2009 to 2018. Replications were high powered (median 99.6% power to detect the original effect size), used the original materials where available, and were peer-reviewed in advance through a standardised protocol. The headline finding is that 55% of claims and 49% of papers replicated in the original direction at conventional significance. The replication rate varies modestly across disciplines, from 43% to 63%, with substantial uncertainty around several of those subfield estimates.

The first thing to say about these numbers is that they should not be read as a verdict on social science overall. The studies sampled are positive results, which is the right thing to test (the file drawer is full of null results that nobody replicates), but it means that some attenuation is mechanically expected through regression to the mean and the limited power of the original studies relative to the replications. Overall, the results suggest a substantial fraction of the empirical literature does not survive a well-resourced replication attempt, that the size of replicable effects is on average about half of what the originals reported, and that this picture is broadly consistent across the social and behavioural sciences rather than being a problem of any one field.

The implications for our own work are several. First, when you read a published finding, your prior should be that the headline effect size is somewhat overstated and that the probability of replication is well short of one. Second, when you do your own work, the practices Miguel et al. (2014) describe are the response Tyner et al. (2026) would recommend: pre-registration to reduce the drift from exploratory to confirmatory framing, open materials so that the next replication does not have to start from scratch, and explicit reporting of effect sizes with uncertainty rather than the binary significant-or-not framing that exaggerates fragile findings. Third, the replicability conversation has moved past the early shock and into the harder territory of asking which design choices, fields, and topics produce replicable findings. This article begins to answer that question.

Reading questions

Honours

7. The headline replication rate in Tyner et al. (2026) is 55%. What do you think this number does and does not tell us about the credibility of social and behavioural science?
8. Effect sizes shrank substantially between the original studies and the replications. What are the most plausible explanations for this pattern, and what are the implications for how we report effect sizes in our own work?

MA/PhD

7. Tyner et al. (2026) provide thirteen alternative estimates of replication success, ranging from 29% to 75%. How should the field choose among these criteria, and what does the spread tell us about the very concept of "replication success"?
8. Compare the replicability findings here with the open-science recommendations from Miguel et al. (2014). Are the practices that Miguel et al. (2014) recommend likely to raise the replicability rate? What additional practices might be needed?

Bazzi and Blattman (2014), "Economic Shocks and Conflict"

Bazzi and Blattman (2014) is the applied article, and it builds on Miguel, Satyanath, and Sergenti (2004), which we read in Week 8. The Miguel et al. (2004) paper instrumented for income with rainfall in Sub-Saharan Africa and found that negative income shocks raised the probability of civil conflict onset, a result that became one of the most influential in the political economy of conflict. Bazzi and Blattman (2014) revisit the same broad question with new data on export price shocks, a different (and arguably more general) instrument, and an explicit and extensive engagement with the limitations of the original design. Their finding is more nuanced and, in important respects, contradicts the Miguel et al. (2004) conclusion. Price shocks have no effect on new conflict, even large shocks in high-risk countries. Rising prices are weakly associated with shorter and less deadly continuations of existing conflicts, which the authors interpret as supporting the state-capacity story (more revenue, better counterinsurgency) rather than the state-capture story (more revenue, more incentive to seize the state).

The methodological lessons are myriad. First, the apparent robustness of the original Miguel et al. (2004) finding depends on choices about sample, controls, instrument, and outcome definition that, when relaxed or generalised, change the substantive conclusion. This is exactly the Plümper and Traunmüller (2020) point about sensitivity, and it is the Clarke (2005) point about the bias structure of regression with controls. The fragility of the result was not visible from the original paper; it took a follow-on paper with new data and a new identification strategy to reveal it. Second, conflict onset and conflict continuation are different processes, and pooling them obscures rather than reveals the dynamics that matter substantively. This is a measurement and operationalisation point that connects back to the conceptualisation readings from Weeks 3 and 4. Third, the way Bazzi and Blattman (2014) communicate their replication-and-extension is itself a model of transparent reporting: they describe what the original found, what they did differently, why those differences matter, and what the substantive conclusion is once the more general design is applied. This is the kind of paper Tyner et al. (2026) would point to as showing how the field should handle the slow accumulation of evidence on a contested question.

There is a fourth lesson relates to how the Miguel et al. (2004) finding influenced policy. Aid agencies, militaries, and stabilisation programmes built strategies around the income-conflict

link. The Bazzi and Blattman (2014) result does not exactly overturn the original article (the contexts and instruments differ enough that the disagreement is partial), but it does substantially weaken the case for the policies that the original supported. This is the human stakes of robustness work. Sensitivity analysis and replication are not housekeeping; they are the part of the discipline that decides which empirical claims will be allowed to drive practice.

Reading questions

Honours

9. In your own words, what does Bazzi and Blattman (2014) change about how we should read the original Miguel et al. (2004) result? Is the original result wrong, or are the two papers addressing slightly different questions?
10. Bazzi and Blattman (2014) distinguish conflict onset from conflict continuation. Why does this distinction matter for inference, and what does it suggest about how we should operationalise dependent variables in conflict research?

MA/PhD

9. The Bazzi and Blattman (2014) extension required new data, a new instrument, and several years of work. What does the cost structure of serious robustness work imply about the incentives in academic publishing?
10. The original Miguel et al. (2004) paper influenced policy. What are the obligations of researchers when their findings are taken up by policy actors, and how should those obligations interact with an evolving evidence base?

Cross-reading question

The five readings this week describe responses at different levels: better sensitivity tests (Plümper and Traunmüller), better measurement of latent constructs (Hollyer, Rosendorff, and Vreeland), institutional practices for transparency (Miguel et al.), evidence on whether those practices have moved the dial (Tyner et al.), and a worked example of patient replication-and-extension (Bazzi and Blattman). If you had to allocate your scarce attention across these levels in your own career as an empirical researcher, which would you prioritise, and why? Is there a level that is missing from this list?

PART 3: CLASS DISCUSSION (DRAWING THE SEMESTER TOGETHER)

This week is the last substantive class before presentations. Rather than run a structured group activity, we are going to spend the rest of today in open class discussion, building on the threats-and-responses framing of the past two weeks and stepping back to review the semester. The goal is to leave today with a clearer sense of how the readings, the design strategies, and the diagnostic vocabulary fit together into a working approach to empirical research, and to have begun the transition into the kind of conversation we will be having in Weeks 11 and 12.

This week's readings

The five articles this week are deliberately cross-cutting. Plümper and Traunmüller (2020) give us reasons to be cautious about the routine sensitivity-test workflow; Miguel et al. (2014) give us a positive programme of practices that pre-empts some of the concerns Plümper and Traunmüller raise; Tyner et al. (2026) give us recent evidence on how far that programme has taken us; Hollyer, Rosendorff, and Vreeland (2014) remind us that all of this presupposes good measurement, which is a problem one layer further upstream; and Bazzi and Blattman (2014)

is the example of what a serious response looks like when the original design turns out to have hidden fragilities. What is missing from the response toolkit that should be there? If you had to write a paragraph about how you intend to handle robustness in your own paper, what would it say after reading these five articles? What practice from the readings would you adopt now, and what practice would you want to adopt but cannot yet?

The semester

This class has moved through five blocks: concepts and measurement (Weeks 3 and 4), case selection and scope (Weeks 5 and 6), causal inference (Weeks 7 and 8), diagnosing threats (Week 9), and responding to them (Week 10). Concepts and measurement are upstream of everything else, because a careful design of an unmeasurable concept does not produce anything. Case selection and scope determine what kinds of inference will be available before any analytical research is conducted. Causal inference is the analytical apparatus we use within the design choices we have made. The threats-and-responses block is what we do when the apparatus runs into the world.

I would like each of you to be able to tell a story about your own project that walks through all five blocks. What is the concept you are studying, and how have you operationalised it? Which cases did you choose, and what does that choice imply about what you can and cannot infer? Which design strategy did you adopt, and what counterfactual does it construct? Which threats are most damaging to your inference, and which are manageable? What practices, ranging from sensitivity tests to pre-registration to honest limitation sections, are you using to respond?

Looking back across the term, where do you think the design tools we have discussed are weakest? Which kinds of questions do they handle well, and which kinds do they not? Most of you are about to go and using these tools in your thesis or dissertation. I would rather you be honest about their limits at the outset than discover those limits too late.

The presentations

The mechanics around the presentations are in the Week 11 and 12 notes and in my presenting guide, but a few points are worth reiterating. The opening should state the question and the headline takeaways in the first ninety seconds, not on the last slide. Limitations belong in the talk, but as a brief, prioritised list rather than a defensive catalogue; the diagnostic discipline from Week 9 and the response toolkit from this week are the kind of tools you chould be drawing on for that section. Discussant comments should be specific, constructive, and brief; the discussant who restates the paper in the first minute has wasted everybody's time. Audience questions should be concrete and focused.

A note on participation during Weeks 11 & 12

This is the kind of session in which the quality of the discussion is set by the audience, not by me. I would like at least one substantive contribution from each of you over the course of the session.

PART 4: WRAPPING UP THE SEMESTER

A few takeaways as we wrap up this semester's substantive discussion. First, the credibility of an empirical claim is set, mostly, before any regression is run or case analysed. The decisions that matter most are about concept, measurement, case selection, and design. The research

design is consequential, but it is downstream of the decisions that determine what it can achieve.

Second, threats to inference are universal; responses are local. Every empirical study has threats, and pretending otherwise is the kind of methodological theatre we have been working against since Week 9. But the right response depends on the question, the data, and the field, and there is no checklist that absolves the student of the work of choosing. The diagnostic vocabulary from Week 9 and the response toolkit from this week are useful when applied with judgment, and the today is part of building that judgment.

Third, transparency is the practice that ties the rest together: pre-registration, open data, honest reporting of effect sizes and uncertainty, and careful description of the limitations the design did not solve. These are not extras to be done if there is time. They are the practices that allow the field to accumulate knowledge rather than recycle confidently asserted findings that do not survive the next attempt to reproduce them. Tyner et al. (2026) is the cleanest available statement of why this matters, and Miguel et al. (2014) is the most concrete available statement of how to do it.

Fourth, the work this semester has been preparation for the work most of you are about to do. The research design papers you are writing, the presentations next week and the week after, and the careers some of you are starting in this seminar all rely on the same disposition: take the question seriously, choose the design that gives you the most credible counterfactual the data allow, diagnose the threats honestly, respond to them in proportion to their severity, and let the reader see what you did. If you leave this semester with that disposition, I will be a happy camper.

What to bring to the next two weeks

Presenters: a one-page research design summary, circulated 48 hours before your session to at least your discussant, with two or three specific questions on which you most want feedback. Slides emailed to me at least an hour before class.

Discussants: a written version of your comments, sent to the presenter and to me within 48 hours of the session.

Audience: each presenter's one-pager read in advance, and at least one substantive question over the two weeks.

I am looking forward to the next two weeks. The work the class has produced so far has been substantively interesting and methodologically careful, and the presentations and the discussant exchanges should be the strongest part of the seminar. See you next week!