Research Design in Political Science
POLS4011/POLS8058

Richard Frank
17 March 2026

WEEK 4: CONCEPTS UNDER PRESSURE, PART 2

PART 1: OVERVIEW

This is the second week of our two-week block on concepts and measurement. Last week we learned what concepts are, how they are structured, what distinguishes a well-formed concept from a vague one, and what happens when conceptual boundaries are unclear. This week we focus on how we get from an abstract concept to something we can observe and measure. The move from concept to measurement is where many research designs quietly fail. We can have a clear concept of state capacity or democracy and still produce a deeply flawed study if the indicators we choose do not actually capture our concept.

"Concepts under pressure" remains relevant this week because operationalisation is where concepts face their hardest empirical tests. It is one thing to define democracy as a system of government that meets certain criteria of contestation, participation, and accountability; it is another to decide whether a country with free elections and a non-independent judiciary counts as a democracy. The concept does not answer the question for us; our operationalisation does. And different operationalisations will give you different answers, which means different datasets, different findings, and potentially different conclusions about the same underlying concept.

Last week's readings gave us a way of thinking about concept structure. Adcock and Collier (2001) give us a four-level framework (background concept->systematised concept->indicators->scores). Goertz (2006, Ch1) shows us how the logical relationships between levels matter (necessary-and-sufficient [AND/OR] versus family resemblance). Collier and Levitsky (1997) shows what happens when a core concept spreads uncontrollably. This week's readings pick up where these readings leave off. Goertz (2006, Ch2) examines whether the way a concept is measured corresponds to how it is defined. Munck and Verkuilen (2002) compare major democracy indices against explicit criteria for conceptualisation, measurement, and aggregation. And Coppedge et al. (2011) present an approach (that would eventually turn into V-Dem) to build a measurement architecture from the ground up that takes (they say) all of these concerns seriously.

All of you should have completed your research design memo by last Friday, so by now, each of you has named the concepts you plan to explore. This week I want you to think about how you will know those concepts when you see them. What will you measure, how will you measure it, and how confident are you that your measures capture what your concepts require? If you cannot answer these questions clearly, you have a measurement problem, and as this week's readings suggest, measurement problems are often conceptual problems in disguise.

Also, now that your research design memos are done and dusted, this is a useful time of the semester to take a moment to reflect. In my experience, several common patterns typically emerge in early research designs: (1) concepts are named but not defined, (2) definitions are borrowed from a single source without considering alternatives, and (3) measures are chosen because they are available rather than because they are valid and reliable. None of these patterns represent fatal errors. They are normal at this stage, but hopefully this week's readings and discussion will give you the vocabulary to diagnose and fix your design's weaknesses as you work toward the final research design paper.

PART 2: READINGS

This week's readings are:

1. Goertz (2006), *Social Science Concepts: A User's Guide*, Chapter 2. Princeton, NJ: Princeton University Press: 27-67.
2. Munck and Verkuilen (2002), "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices," *Comparative Political Studies* 35(1): 5–34.
3. Coppedge et al. (2011), "Conceptualizing and Measuring Democracy: A New Approach," *Perspectives on Politics* 9(2): 247–267.

1. Goertz (2006), Chapter 2

Chapter 1 of Goertz (2006) described concepts as having a three-level structure: the basic level, secondary dimensions, and indicator/data level. Chapter 2 turns to the consistency problem: does the operational measure behave in ways that are consistent with the theoretical definition? This sounds straightforward, but Goertz (2006) shows that it is surprisingly difficult to answer and that many widely used measures in political science fail this test.

The core idea is deceptively simple: if your concept says Y has three necessary dimensions (A, B, and C), then your measure should require all three. If your concept says the dimensions relate by family resemblance (any two of three suffice), your measure should reflect that. In practice, many measures violate this consistency requirement. Researchers define a concept one way and then measure it another way, sometimes because data availability forces compromises, sometimes because the gap goes unnoticed. Goertz's (2006) argument is that this is not just a technical problem but an inferential one. If your measure does not match your concept, your findings may be about something other than what you think they are about.

Goertz (2006) continues to stress the difference between necessary-and-sufficient versus family-resemblance concept structures because the aggregation rules you use to combine indicators should reflect which structure your concept has. If democracy requires *both* contestation and participation (necessary-and-sufficient), then averaging scores across those dimensions is inappropriate. A country scoring very high on contestation but zero on participation is not "moderately democratic." It fails a necessary condition. However, averaging is exactly what many indices do (I should know as I sadly helped generate a few in my time), which means the maths contradicts the conceptual logic. The minimum function (take the lowest score across necessary dimensions) would be the appropriate aggregation rule for necessary-condition concepts. For family resemblance concepts, where sufficiency on any dimension can compensate for weakness on another, averaging or additive rules make more sense.

Most of you will not (thankfully) be building complex indices, but the logic applies to any research design. If you argue that your independent variable has two dimensions that are both necessary, and then you construct a single measure by averaging them, you introduce an inconsistency between your theory and your measurement. Goertz (2006) is teaching us to ask does my measure do what my concept says it should do? This is a question that should be answerable, and this chapter provides the tools to answer it. He also helpfully summarises the implications of his argument for researchers like us at the end of each section.

Reading questions

Honours students

1. What is the difference between a necessary-and-sufficient concept structure and a family-resemblance structure? Can you think of a concept from your own research area that clearly fits one or the other?
2. Goertz (2006) argues that averaging scores across dimensions is inappropriate for necessary-condition concepts. Why? What would be a concrete example where averaging produces a misleading result?

MA/PhD students

1. How does the concept–measure consistency problem relate to Adcock and Collier's (2001) four-level framework from last week? At which level(s) of their framework does the inconsistency Goertz (2006) highlights typically enter?
2. Consider one of the major variables in your own research. What is its internal structure? Are the dimensions related by necessity, sufficiency, or something else? Does the standard operationalisation in your subfield reflect that structure, or does it default to additive/averaging aggregation regardless?

2. Munck and Verkuilen (2002)

Munck and Verkuilen (2002) take the general problem Goertz (2006) raises (does the measure match the concept?) and apply it systematically to arguably the most important/consequential measurement process in comparative politics: measuring democracy. They evaluate nine major democracy indices (including Freedom House, Polity, Vanhanen, Bollen, and others) against a common framework organised around three challenges: conceptualisation, measurement, and aggregation, challenges that should be becoming familiar by now. Their findings are sobering if unsurprising by now. Every major index has significant weaknesses, and the weaknesses are often at the level of conceptualisation.

This paper is organised around three sequential tasks that any measurement process involves. First, *conceptualisation*: identifying the relevant attributes of the concept and organising them logically. Here the risks include maximalist definitions (too many attributes, which leads to unmanageable complexity), minimalist definitions (too few attributes, which misses essential content), and redundancy (including attributes that overlap). Second, *measurement*: selecting indicators for each attribute and devising coding rules. The risks include using indicators that do not clearly map onto the specified attributes, coding rules that are too vague to be applied consistently, and insufficient attention to intercoder reliability. Third, *aggregation*: combining indicators into an overall score. This is where Goertz's (2006) concern about aggregation rules becomes concrete. Munck and Verkuilen (2002) show that many indices aggregate in ways that are mathematically convenient but conceptually unjustified.

In their analysis, no index performs well across all three challenges. Freedom House (FH) receives probably the harshest critiques. For example, they argue that FH has broad empirical coverage and a long time series, but its conceptualisation is unclear (the relationship between political rights and civil liberties to democracy is underspecified), its coding rules are opaque, and its aggregation method is not well justified. By contrast, Polity has a clearer conceptual foundation focused on contestation and executive constraints, but it excludes participation entirely, which is a major conceptual choice that many users (myself included up until now) do not notice. Vanhanen's (2000) index is transparent and replicable (it uses only two clear indicators: voter turnout and vote share), but its conceptualisation is thin gruel, and the indicators are crude proxies for what most academics mean by democracy. The overall lesson here is that measurement choices involve trade-offs, and the problem is not that trade-offs exist but that they are often invisible to users who download a dataset and treat a democracy score as an unproblematic fact (as I used to do).

Munck and Verkuilen (2002) use democracy as their focus, but the three-challenges framework is portable to other concepts, including those that you are focusing on. Any time you or I use an existing index or scale (corruption indices, state fragility indices, press freedom scores, electoral integrity measure, etc) we should be asking the same questions: how is the concept defined, what indicators are used, and how are they combined? This article teaches us a general habit of critical evaluation that applies to any measure we might come across.

Reading questions

Honours students

3. Pick one of the democracy indices discussed in the paper (Freedom House, Polity, or Vanhanen). What is its main conceptual weakness according to Munck and Verkuilen (2002), and why does that weakness matter for empirical research?
4. Munck and Verkuilen (2002) argue that many indices aggregate indicators in ways that are not justified by the underlying concept. What does this mean in practical terms? Can you construct a simple example where the aggregation rule changes which country appears "more democratic"?

MA/PhD students

3. Apply Munck and Verkuilen's (2002) three-challenges framework to a measure you are using or plan to use in your own research. How well does it perform on conceptualisation, measurement, and aggregation? Where are its vulnerabilities?
4. The article implies that there is a tension between parsimony and validity in index construction. Is it possible to resolve this tension, or is it inherent in any attempt to reduce a multidimensional concept to a single score? What are the implications for researchers who need a single number for a regression?

3. Coppedge et al. (2011)

If Munck and Verkuilen (2002) find a bunch of problems with existing democracy measures, Coppedge et al. (2011) propose to build something better (and then actually does it). This article is the first description of what would become the Varieties of Democracy (V-Dem) project. It proposes a measurement structure for democracy that is explicitly designed to address the weaknesses the literature has identified: conceptual disagreement about what democracy means, aggregation choices that smuggle in unexamined theoretical assumptions, and coding procedures that lack transparency and reliability checks.

V-Dem's big idea is to disaggregate rather than define democracy as a single thing. The project identifies multiple distinct "varieties" or "principles" of democracy (electoral, liberal, participatory, deliberative, and egalitarian) each with its own conceptual logic. Rather than asking "how democratic is country X?" as a single question, V-Dem asks "how electoral is country X? How liberal? How participatory?" and so on. Each principle is then built up from fine-grained indicators at the lowest level, aggregated through explicit rules to the principle level, and only combined into an overall score as a final step. Even then, V-Dem provides both the disaggregated scores and the aggregate index, so users can choose the level of analysis appropriate to their research question. Although to be honest most articles that use V-Dem's democracy measures stick to their overall score. Plus ça change…

This article reads like a direct response to the problems raised by this week's (and last week's) readings. In response to Collier and Levitsky's (1997) observation that democracy-with-adjectives proliferates because scholars disagree about what democracy means, V-Dem accommodates multiple conceptions rather than choosing one. In response to Goertz's (2006) concern about concept–measure consistency, V-Dem makes the aggregation rules explicit and theoretically justified at each level. In response to Munck and Verkuilen's (2002) critique that coding procedures are often opaque, V-Dem will use multiple independent coders per country-year and applies a (slightly overly fancy in my opinion) measurement model to estimate coder reliability and adjust scores accordingly. V-Dem's approach to coding is distinctive and worth discussing. Rather than relying on a small team of coders using fixed rules (the Freedom House model), V-Dem recruits multiple country experts for each country-year observation. The project then uses an item response theory (IRT) model to estimate each coder's reliability and to produce final estimates that account for coder disagreement. This is an attempt to address the intercoder reliability problem that Munck and Verkuilen flagged.

V-Dem is not without its own vulnerabilities. The project's scope is enormous (hundreds of indicators, thousands of coders, more than two centuries of coverage), which introduces complexity that creates its own risks. The measurement model requires assumptions about coder behaviour that may not always hold. And the sheer number of available indicators and indices can be overwhelming for users (at least speaking for myself), leading to the same black-box problem Munck and Verkuilen (2002) identified. Researchers may download V-Dem data without fully understanding how their indices are constructed, which simply shifts the opacity to a different level.

Reading questions

Honours students

5. V-Dem identifies six "principles" of democracy (electoral, liberal, participatory, majoritarian, deliberative, and egalitarian). Choose two and explain how they differ. What would a country look like that scores high on one but low on the other?
6. Coppedge et al. (2011) argue that disaggregation is better than choosing a single definition of democracy. But is there a cost to not choosing? When might a scholar need a single definition, and how should they decide which one to use?

MA/PhD students

5. Evaluate V-Dem using Munck and Verkuilen's (2002) three-challenges framework. How well does V-Dem perform on conceptualisation, measurement, and aggregation

compared to the older indices Munck and Verkuilen (2002) reviewed? Where does it improve, and where might it still fall short?
6. V-Dem uses an item response theory (IRT) model to aggregate expert ratings. What assumptions does this require about coder behaviour, and what happens to the estimates if those assumptions are violated? How does this compare to simpler aggregation methods like averaging?

Overall reading questions for all students

1. Goertz (2006) argues that the aggregation rule should follow from the concept structure. Munck and Verkuilen (2002) show that many indices violate this principle. Does V-Dem succeed where the older indices failed, or does it introduce new aggregation problems of its own?
2. Trace a single concept, democracy, across all six readings in this two-week block. How has the way we think about defining and measuring democracy changed from Collier and Levitsky (1997) (naming the problem) to Adcock and Collier (2001) (building a framework for evaluation) to Goertz (2006) (specifying structural consistency) to Munck and Verkuilen (2002) (systematic evaluation) to Coppedge et al. (2011) (attempted solution)? What problems remain?
3. Think about a concept in your own research. If someone applied Munck and Verkuilen's (2002) three-challenges framework to your planned operationalisation, where would it be most vulnerable? What could you do to strengthen it?
4. Cast your mind back to Hyde (2007) from Week 2. She needed an operational measure of "election fraud" to make her natural experiment work. How would you evaluate her operationalisation using this week's tools? Where is it strong, and where might it be challenged?

PART 3: GROUP WORK

This week's group activity is a single, shared exercise that all students complete simultaneously in groups of 2-4 students. The goal is to give you hands-on practice with the focus of this week's readings, identifying the gap between how a concept is defined and how it is measured, and then apply it directly to your own research.

The Measurement Trade-Off Card (15–20 minutes)

Work in groups of 2-4 students. Each group picks a concept relevant to one member's research (ideally from their research design memo) and fills out a simple three-row "trade-off card" on a sheet of paper or in a shared document. If you have the time, you can fill out cards for more than one student's project.

Your card should have three rows:

1. **Row 1: What does the concept require?** Write the concept name, a one- or two-sentence definition, and whether its dimensions are related by necessity (all must be present) or family resemblance (some subset suffices). Draw on Goertz's (2006) framework here.

2. **Row 2: How would you (or how does someone) measure it?** Name a specific indicator or existing index you would use (or that is commonly used in the literature).

Be concrete: not “survey data” but “Freedom House political rights score” or “count of protest events in ACLED.”

3. **Row 3: Where is the gap, and what trade-off are you making?** In one or two sentences, identify the biggest concept–measure inconsistency. What does the measure capture that the concept does not require, or what does the concept require that the measure misses? Use Munck and Verkuilen’s (2002) three challenges as a diagnostic guide.

The card format keeps us brief and concise while maintaining the focus on the core analytical shift from this week’s readings: identifying the gap between how a concept is defined and how it is operationalised.

All together debrief (10 minutes)

Each group should read your Row 3 from your card and briefly answer the following questions:

1. Which of these trade-offs is most consequential for the conclusions you might reach using your approach?

2. Which gaps could be fixed with better data, and which are inherent in the concept itself?

3. Do you see the same kind of concept–measure inconsistency that Munck and Verkuilen (2002) found in the democracy indices showing up in your own work? Why/Why not?

4. What is the single most important measurement decision you still need to make for your own project, and what framework from this two-week block will help you make it?

PART 4: WRAPPING UP AND LOOKING AHEAD

This is a natural moment to take stock. Over the past two weeks, you have developed a toolkit for thinking about concepts and measurement that includes Adcock and Collier’s (2001) four-level framework, Goertz’s (2006) structural analysis, Collier and Levitsky’s (1997) warning about conceptual proliferation, Munck and Verkuilen’s (2002) systematic evaluation framework, and V-Dem as a case study of how to do measurement self-consciously. The message is not that perfect measurement is achievable, but that thoughtful measurement is, and thoughtful measurement requires making your conceptual and operational choices explicit, justified, and open to scrutiny.

The next assessment, the critical review asks you to select a peer-reviewed article and evaluate its research design. You should now have the tools to assess not just whether an article’s methods are appropriate in some general sense, but specifically whether the article’s operationalisations are valid: whether the concepts are well-defined, whether the measures match the definitions, and whether the aggregation (if any) is justified. I encourage you to start thinking about which article you will review and to look specifically at the article’s measurement choices as a potential focus for critique.

Weeks 5 and 6 are the next two-week block, which shifts from what we are studying (concepts and measures) to where and how many (case selection and scope). The logic connects directly:

once you know what you are measuring, the next research design question is which cases to observe, how many you need, and what you can claim based on the cases you choose. Those of you who have worked through the concepts block carefully will find that case selection decisions are easier to make (and easier to justify) when the underlying concepts and measures are clear. Those who have not will find that vague concepts produce vague case selection rationales.