Think Locally, Regress Globally: Making the Most of Conventional IR Data

Carlos Felipe Balcazar and Matt Malis

May 25, 2021

1 Introduction

Recent years have seen an incredible proliferation of "conventional" datasets for the study of international relations (IR)—datasets that report a measure of some political, social, or economic phenomenon occurring in some period of time (typically a year) at some aggregated geographical level (typically a sovereign state, or a state-to-state dyad). Not only the data themselves, but the software for organizing, preparing and analyzing the data are increasingly accessible for well-resourced institutions and independent researchers alike. It is now easier than ever for a researcher, armed with only an internet connection and a modest amount of computing power, to conduct sophisticated statistical analyses of international interactions.

In light of these developments, this essay seeks to provide practical guidance for applied quantitative IR researchers regarding the steps of the research process in between theory development and statistical analysis. That is, given a clearly articulated theoretical prediction, what must be done before the researcher can run a regression? This chapter primarily addresses decisions pertaining to the selection of a sample of analysis, and the selection of variables to operationalize theoretical quantities of interests, with a focus on the implications of these decisions for internal and external validity and statistical power. Treatments of both earlier stages (e.g. theory development) and later stages (e.g. model specification) of the research process can be found in chapters by Barnum (2021), Butler (2021), Chiozza (2021), Gonzalez and Poast (2021), and Ritter (2021) within this volume.

Our overarching claim is that the increasing accessibility of conventional IR data, and tools for analyzing these data, require that researchers be ever more careful about how they approach their analyses. We suggest that IR researchers develop the practice of "thinking like an experimentalist": treat your data as if it were costly to obtain, with outcomes unknown ex-ante, and invest the time upfront to think through the myriad decisions which will have to be made throughout the process of analysis.

We frame our discussion through the device of a pre-analysis plan (PAP) of the kind used frequently in experimental research, but very rarely in observational studies. The primary function of a PAP is generally considered to be as a sort of hands-tying commitment device—a "Ulysses Pact", as described by Janzen and Michler (2021)—which, in combination with a pre-registration process, limits a researcher's latitude for "fishing", or selectively reporting analyses on the basis of the results they produce. Our focus is on a second, less-appreciated function that PAPs can serve even when their hands-tying value is limited: forcing the researcher to confront the challenges and complications that inevitably arise in the analysis stage, and to develop principled solutions to those problems before seeing how those solutions bear on the statistical results. We discuss these issues theoretically and provide references to a range of resources to help readers apply these concepts in practice.

2 Running example: Fiscal windfalls and governance quality

To ground our discussion, we consider as a running example the following research question: What is the impact of fiscal windfalls on the quality of governance? We use the term "fiscal windfalls" to refer generally to resources a government accrues through means other than taxation ("easy money" as characterized by Bueno de Mesquita and Smith (2013)). We phrase the research question broadly to allow for a range of operationalizations of both the independent and dependent variables, as well as a number of reasonable options regarding the particular empirical context that the researcher could choose to analyze.

An attractive feature of this research question is that, for any given operationalization or empirical context, one could reasonably theorize the effect of fiscal windfalls working in either direction, either to the improvement or detriment of governance quality. Standard arguments predicting a negative effect highlight the incentives of incumbents—windfalls make those in power less accountable to the tax-paying public, and more willing to employ repression and corruption to extract and secure rents from office-holding—as well as incentives of challengers and insurgents—those seeking to undermine the established order anticipate a greater reward from doing so as a result of windfalls. In contrast, arguments predicting a positive effect point to the enhanced ability of the government to buy off would-be challengers (and relatedly, an opportunity-cost effect which disincentivizes anti-government mobilization), or alternatively to the long-run impact of income on democracy.

A sample of papers that examine some version of this general research question can be found in Table 1. These papers come from a variety of journals covering international relations, comparative politics, and economics, illustrating the diversity of possible approaches to the topic. The subject of fiscal windfalls is of particular interest to scholars of international relations because the sources of these windfalls often originate outside a country's borders, for instance in the form of foreign aid, or of price shocks that propagate through the global economy. But as is also evident from this sample of papers, similar theoretical arguments can be applied to the study of domestically-originating windfalls as well. We return to the question of what makes a political question distinctly "international", and what are the relative merits of cross-country versus within-country research designs, in Section 4 of this chapter.

Table 1: Empirical studies of fiscal windfalls and governance

Author & Year	**Unit of observation**	**Cross-sectional coverage**	**Time period coverage**	**Measure of windfall**	**Measure of governance**
Goldberg, Wibbels and Mvukiyehe (2008)	State-year	All US States	1929-2002	Natural resources as share of state GDP	Margin of victory in gubernatorial elections, and vote share of the incumbent governor
Djankov, Montalvo and Reynal-Querol (2008)	Country × 5-year period	108 aid-recipient countries	1960-1999	ODA as share of GDP	Change in "checks" variable from Database of Political Institutions, and (separately) change in "democracy" measure from Polity
Ramsay (2011)	Country-year	48 oil-producing countries	1968-2002	Value of oil production	Polity2 score from Polity IV
Nielsen et al. (2011)	Country-year	139 countries	1981-2005	Aid shocks, using ODA from AidData	Armed conflict onset, from UCDP/PRIO
Brückner, Ciccone and Tesei (2012)	Country-year	170 countries	1960-2007	International oil prices	Polity2 score, constraints on the executive, and political competition measures from Polity IV
Dube and Vargas (2013)	Municipality-year	978 municipal units	1988-2005	International commodity price shocks	Guerrilla attacks, paramilitary attacks, clashes or casualties in a given municipality
Brollo et al. (2013)	Municipality	2,877 Brazilian municipalities	2001-2009 (pooled for cross-sectional analysis)	Federal transfers to municipal governments	(i) Incidence of corruption by incumbent mayors; (ii) Mayoral candidate quality (measured by education)
Nunn and Qian (2014)	Country-year	125 non-OECD countries	1971-2006	Amount of wheat aid shipped to a recipient country from the US	Occurrence of conflict, from UCDP/PRIO

Table 1 (Cont'd)

Author & Year	**Unit of observation**	**Cross-sectional coverage**	**Time period coverage**	**Measure of windfall**	**Measure of governance**
Crost, Felter and Johnston (2014)	Municipality	222 eligible or nearly-eligible municipalities	2002-2006 (pooled for cross-sectional analysis)	Aid provided through a community-driven development program, funded by the World Bank	Conflict casualties, disaggregated by insurgents, government forces, and civilians
Dube and Naidu (2015)	Municipality-year	936 municipalities in Colombia	1988-2005	US military and antinarcotics aid to Colombia	Conflict-related incidents (paramilitary, government or guerrilla attacks)
Caselli and Tesei (2016)	Country-year	131 countries	1962-2009	3-year rolling average change in global price of primary export commodity	Year-to-year change in Polity2 score
Sexton (2016)	District-week	398 districts in Afghanistan	May 2008 to Dec. 2010 (138 weeks)	Spending on civilian aid projects through the Commander's Emergency Response Program (CERP)	Geocoded incidents of violence by both pro- and anti-government forces
Carnegie and Marinov (2017)	Country-year	115 former colonies	1987-2006	ODA from the EU	CIRI human empowerment index; and Polity2
Berset and Schelker (2020)	Municipality-year	162 Swiss municipalities	2008-2016	Tax revenue shock from IPO of Swiss firm on London Stock Exchange	Municipal fiscal revenue, expenditures targeting specific groups, and user charge hikes
Balcazar and Ch (2021)	Country-year	54 countries	1870-1905	Tariff revenues	5-year average for various V-Dem scores

3 Thinking through a pre-analysis plan

To organize our discussion of the steps of the research process that occur between theory development and statistical analysis in a quantitative IR study, we consider how a researcher would think through creating a pre-analysis plan (PAP) for her study. As described by the Evidence in Governance and Politics (EGAP) project, a PAP is "a document that formalizes and declares the design and analysis plan for your study. It is written before the analysis is conducted and is generally registered on a third-party website."[1]

The primary function of a PAP in the social sciences is typically thought to be as a means for the researcher to commit herself to a course of analysis ex-ante, and to avoid developing hypotheses or selectively reporting results of hypothesis tests ex-post (practices variously referred to as "fishing", "p-hacking", or "HARKing" (Hypothesizing After Results are Known)).[2] As Janzen and Michler (2021) discuss, this logic reflects the literary trope of "Ulysses' Pact": much like Homer's protagonist had to tie his hands to a mast to prevent himself from steering his ship off course, the researcher must find a way to resist the Siren song of selectively reporting statistically significant results that confirm her theoretical predictions (or revising those predictions to fit the statistical results). Thus a PAP, in combination with a pre-registration process which makes the PAP publicly accessible, is conventionally seen as a means of enhancing the transparency and thus the credibility of published scientific research.

Seen in this light, PAPs would seem to be of little use for observational research, which constitutes the overwhelming majority of quantitative IR scholarship. Since observational data are generally available at the time the PAP would be written, the researcher can no better commit to refrain from fishing or HARKing in the PAP than she could in the final write-up. However, the PAP's utility as a commitment device is only of its potential virtues. As Janzen and Michler (2021) write:

> Although pre-analysis plans are usually promoted for the benefits they provide to the profession, there are additional practical benefits for a researcher. First, it is simply a good exercise to carefully think through how the data will be used before collecting or acquiring access to the data. In our experience, the attention to detail and thoughtfulness required to write a good pre-analysis plan, leads to an equally careful and thoughtful final analysis. Second, we have found that pre-analysis plans can be particularly helpful when working as part of a research team. ... Third, soliciting feedback at the planning stage can prevent high-cost mistakes, some of which can be impossible to alter ex post, dramatically improving research quality. (p.10)

In a similar vein, Gelman (2013) notes that "Often it is only when focusing on the write-up that we fully engage with our research questions"; he likewise acknowledges the benefits to the researcher of soliciting feedback on a PAP, because "exposing an idea to outside eyes can reveal flaws in the plan."

With these benefits in mind, how might a researcher go about executing a PAP for an observational IR study? There is no single template for a PAP, and no universally-agreed set of questions that a PAP must address. But, as Christensen and Miguel (2020) write, "there appears to be a growing consensus that pre-analysis plans in the social sciences should consider discussing at least the following list of ten issues:

1. study design
2. study sample

[1] `https://egap.org/resource/10-things-to-know-about-pre-analysis-plans/`

[2] For more on this perspective on PAPs, see Humphreys, Sanchez de la Sierra and Van der Windt (2013) and Christensen and Miguel (2020).

3. outcome measures
4. mean effects family groupings
5. multiple hypothesis testing adjustments
6. subgroup analyses
7. direction of effect for one-tailed tests
8. statistical specification and method
9. structural model
10. timestamp for verification

Pre-analysis plans are relatively new to the social sciences, and this list is likely to evolve in the coming years as researchers explore the potential, and possible limitations, of this new tool."

While Christensen and Miguel's analysis focuses on PAPs for experimental research, and particularly on their utility as a hands-tying mechanism, each of the points listed (save for perhaps the last one) is a relevant consideration for the observational researcher who makes no claim of using the PAP as a credible commitment device. For the purposes of the present chapter, we restrict attention (more or less) to the first three points raised above. More specifically, our discussion will be oriented around the following questions:[3]

1. Study sample:
   (a) What is the population of interest for the study?
   (b) What is the sample of analysis?
   (c) How does the sample differ from the population, and how will estimated effects generalize to the population?
2. Variables:
   (a) What are the main variables of interest in your study?
   (b) How will the variables be discretized or otherwise transformed for analysis?
3. Measurement:
   (a) How (in)accurately are your variables measured?
   (b) Is measurement error random or systematically biased? Is it correlated with other variables?
4. Missingness:
   (a) What portion of your observations are missing for each variable?
   (b) What is the mechanism giving rise to missingness?
   (c) Are the observations missing at random, or is missingness correlated other variables?
5. Statistical power:
   (a) What is the variability of your outcome measure, and of your effect size?
   (b) What dependencies exist among your observations?
   (c) What is the minimum effect size you will be able to detect?

We address each of these questions in turn.

[3]These questions also borrow partially from Alejandro Ganimian's Pre-Analysis Plan template, which can be found here: `https://www.bitss.org/resources/pre-analysis-plan-template/`

4 Study sample

The first questions we consider pertain to the population of interest for the study, the sample of analysis, and the relationship between the two. A quantitative IR study typically begins (perhaps implicitly) with the objective of saying something about the entire global population of countries (or dyads or k-ads of countries; see Gonzalez and Poast (2021)), over as long a time period as possible. The researcher proceeds to construct a time-series cross-sectional (TSCS) dataset with "all countries" (or "all k-ads"),[4] in which each observation constitutes a cross-sectional unit (country or k-ad) measured at some point in time (typically a year). As a conceptual matter, statistical inference is conducted by treating this global sample (again, perhaps implicitly) as just one realization of a notional "superpopulation" consisting of infinitely many global samples of country-years.[5] As a practical matter, the sample of analysis is typically thought of as unproblematically representative of the population that the researcher wishes to study.

4.1 Dataset coverage

We wish to point out three distinct complications to this seemingly straightforward relationship between the population of interest and the sample of analysis in a cross-country IR study. First, how exactly should we define a country? An answer to this question is essential to understanding what are the observations that constitute our sample. The Correlates of War Project (2017) defines an entity to be a "state member of the international system" if it satisfies the following criteria: "1) prior to 1920, the entity must have population greater than 500,000 and have had diplomatic missions at or above the rank of charge d'affaires with Britain and France; 2) after 1920, the entity must be a member of the United Nations or League of Nations, or have population greater than 500,000 and receive diplomatic missions from two major powers." One may reasonably question any of these criteria, but they at least provide a clear definition for what is considered to be a sovereign state in the international system, and this definition is shared across the various Correlates of War datasets (see Table 2 below).

Most other datasets are not as explicit in stating the precise criteria for an entity's recognition as a sovereign state and inclusion in the sample. Difficulties may thus arise in merging together data from different sources, which may follow different definitional criteria or different naming conventions for members of the international system. Especially challenging are situations in which the international system changes, for instance in cases of state dissolution, secession, or (re-)unification. Compounding these difficulties is the fact that state recognition is an inherently political process—almost inevitably related, to one degree or another, to the political processes that an IR researcher hopes to study. As one example of the methodological consequences of this political problem, datasets provided by international organizations tend to include states only as they currently exist in the system: the World Bank's World Development Indicators, for instance, code Germany as a unified country dating back to 1970, while Yemen only appears in the data as "Republic of Yemen" beginning in 1990. If the researcher is not careful in reconciling these cases, and perhaps supplementing their data with additional sources, then an important restriction has been applied to the population about which inferences may be drawn from the sample of analysis: rather than saying something about the population of "all countries", the study may only be informative of the population of countries that currently exist, and that have existed as stable countries for some

[4] In some applications, researchers will limit the dataset to "politically relevant" or "politically active" dyads or k-ads. We omit consideration of this decision from the present chapter, but refer readers to Lemke and Reed (2001), Quackenbush (2006), and Gonzalez and Poast (2021) in this volume for further discussion.

[5] This is in the case that the researcher adopts a frequentist framework to statistical inference; for a critique of the applicability of this framework for cross-national research, see Western and Jackman (1994) (or Gill (2001) for a similar critique in the context of state-level data in the US). Abadie et al. (2020) discuss how a frequentist framework can be applied to such data by thinking of uncertainty arising from a hypothetical treatment assignment mechanism, rather than a hypothetical sampling mechanism from an infinite superpopulation.

time, and whose existence is sufficiently politically uncontroversial as to allow their inclusion in a given dataset.

A second complication in relating the sample of analysis to the population of interest follows from a more general issue of data availability and dataset coverage. Separate from any definitional questions of what constitutes a sovereign state, some datasets may simply be missing values for certain countries in certain years. We discuss this problem of missing data in more detail in Section 7 below. A straightforward implication is that, if missingness is related to some other relevant characteristic—countries with low state capacity may be less likely to collect and report economic data, for instance—then the population which the study can inform us about is again limited in an important respect.

4.2 Regression weighting

A third issue is somewhat subtler than the first two, and follows from a technical point regarding the mechanics of multivariate regression. The basic concern is that even if a sample of analysis contains all countries in the international system (setting aside the two issues discussed above), those countries are not contributing evenly to the estimated effects. In other words, the "nominal sample" of countries appearing in the dataset may differ substantially from the "effective sample" which is used to estimate the regression coefficient. We provide here a brief and intuitive discussion of the problem, following the presentation in Aronow and Samii (2016), and encourage readers to review that article for a more thorough treatment.

Consider a sample indexed by $i = 1, \ldots, n$. We can think of each i as representing a single observation (e.g., Peru-1978) or a group of observations (e.g., all years of Peru observations). Suppose that each observation has its own effect, τ_i, of the main "treatment" variable of interest. (We will use the term "treatment" to distinguish the main independent variable in an analysis from the other independent variables in the model, which we will label "covariates", while acknowledging that a typical IR dataset will include no variables which actually represent treatments assigned and manipulated by the researcher.) Returning to our empirical example of fiscal windfalls, we would think of τ_{Peru} as denoting the effect of windfalls on governance quality in Peru. A standard quantitative IR study will estimate a regression equation of the form

$$Governance\,Quality_i = Windfalls_i\,\beta + X_i'\,\gamma + e_i$$

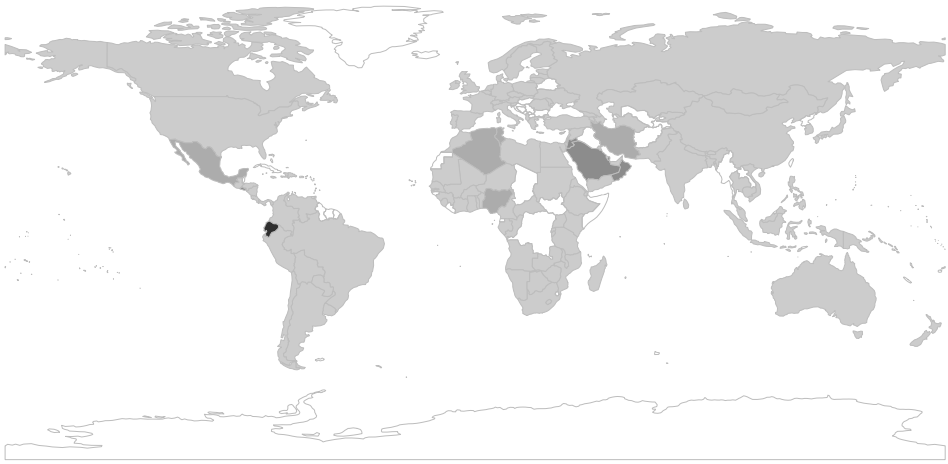
where X_i is a vector of covariates (often including some combination of control variables, lagged dependent and/or independent variables, and unit and/or time fixed effects) and e_i is an error term. The researcher in this case typically aspires (perhaps implicitly) to estimate an average treatment effect, $\tau = E[\tau_i]$, such that the reported results are broadly informative of how fiscal windfalls impact governance in the full population of countries.

What Aronow and Samii (2016) show is that multiple regression—whether a linear regression or a generalized linear model such as logit, probit, Poisson regression, or Cox proportional hazards—generally fails to recover an unbiased estimate of the average effect τ. Rather, the estimated regression coefficient $\hat{\beta}$ converges to a weighted average of unit-specific effects, such that

$$\hat{\beta} \xrightarrow{p} \frac{E[\omega_i \tau_i]}{E[\omega_i]}$$

where $\xrightarrow{p}$ denotes convergence in probability (that is, the value that $\hat{\beta}$ approaches as the sample size gets large). Here ω_i denotes a weight on unit i's effect, where the weights may be quite unevenly distributed across units. In particular, each unit's weight is proportional to the conditional variance of that unit's treatment assignment—that is, how poorly the treatment variable is predicted by the other covariates in the model. If Peru's fiscal windfalls can be accurately predicted by the control

Figure 1: Regression weights in Caselli and Tesei (2016)



Note: Regression weights for the $\Delta Pr_a \times Pl_{t-4,a}$ variable from Model 3 of Table 3 in Caselli and Tesei (2016); computed following the procedure outlined in Aronow and Samii (2016). Darker shading denotes higher weight; white denotes missing values.

variables and fixed effects in the model, then ω_{Peru} will be low, and Peru's treatment effect τ_i will be contributing little to the regression estimate $\hat{\beta}$.

How much does this matter in practice? To illustrate the consequences of uneven weighting in a real-world application, we consider the paper by Caselli and Tesei (2016) that appears in Table 1.[6] The sample consists of 4,745 country-year observations, covering 131 countries and the years 1962-2009. Specifically we examine Model 3 of Table 3 from the paper, where the treatment variable of interest is an interaction between commodity price shocks and Polity score among autocracies (labeled "$\Delta Pr_a \times Pl_{t-4,a}$" in the original table). We follow the procedure outlined in Aronow and Samii (2016) to estimate observation-specific regression weights, and then sum these weights by country. Each country's respective share of the total regression weights is depicted in Figure 1.

What is immediately apparent is that the distribution of regression weights across countries is far from even. The primary contributor to the estimated effect is Ecuador, by a substantial margin, with disproportionate contributions also coming from a handful of Persian Gulf states. While the nominal sample consists of 131 countries, it turns out that 50% of the regression weight comes from ten countries, and 90% of the weight comes from forty countries. In short, the results do not provide nearly the degree of generalizability as one would intuitively expect from an all-inclusive cross-country dataset.

There are of course exceptions to the general inclination, mentioned above, of quantitative IR scholars to analyze datasets that include as many countries (or k-ads) as possible. For some studies, the scope conditions of the theory are explicitly delimited to a certain subset of countries (for instance, oil-exporting countries in Ramsay (2011)). Some IR studies restrict attention to a single country and focus on variation at the subnational level. This could be because the researcher has identified a particular source of exogenous variation she wishes to exploit, which is only applicable to a specific context (Crost, Felter and Johnston, 2014; Sexton, 2016); or because the theoretical

[6] We use this example simply because the replication code is readily accessible and the main empirical specification is a linear model estimated by OLS. The problems of unequal weighting are likely to be similarly if not more pronounced in the other studies included in Table 1, only less straightforward to characterize and visualize.

mechanism operates at a subnational level, and aggregating the data to the national level would obscure the variation of interest (Dube and Naidu, 2015).

Some IR scholars are averse to these sorts of within-country designs on the grounds that they lack generalizability beyond the particular context of analysis. While acknowledging that these designs may enjoy some advantages with regards to the internal validity of the findings, these critics remain concerned over the ability of such designs to speak to the broader political phenomena that IR scholars seek to explain. Yet as revealed in the preceding discussion, an analysis of an all-encompassing cross-national dataset may not yield the sort of generalizable findings that many scholars hope to produce. Put differently, by Aronow and Samii (2016, 251):

> [T]his perceived trade-off [between internal and external validity], and the choice that it suggests, is an illusion... Even though one starts with a sample representative of the population of interest, what one obtains is an effect for which there are still reasons to question generalizability to the population of interest. External validity problems have not been avoided.

5 Variables

The next set of questions we address pertains to the operationalization of key theoretical quantities of interest. Given a sample of analysis, and given a set of theoretical variables that we expect to relate to one another in some way, how do we go about finding and selecting measurements of those variables and preparing them for analysis? We will first note that the sequencing of decisions that we outline here—select a sample of analysis, then choose the variables to analyze—is not always so clear-cut in practice: the ideal measurement of a theoretical variable may have severely limited coverage, and the researcher may opt for a less-perfect measurement that is available for a larger portion of the sample (or vice-versa). In fact, all of the decisions we discuss here will often be made through an iterative process, where earlier decisions may be revisited in light of later realizations. We simply offer this sequence as a reasonable starting point for the development of a research design.

5.1 Data resources

As mentioned at the outset of this chapter, there has been considerable proliferation of cross-national time-series datasets over recent years. For the applied researcher, the plethora of options can be at times overwhelming, leaving her unsure of where to even begin to search for a variable of interest. Fortunately, along with newly available datasets, we have also seen the development of various resources that help researchers organize their data collection efforts. We list several of these resources in Table 2 and briefly discuss each one.

Table 2: Data resources

NewGene (Bennett, Poast and Stam, 2019)	"*NewGene* is a stand-alone Microsoft Windows and OSx-based program for the construction of annual, monthly, and daily data sets for a variety of decision-making units (e.g., countries, leaders, organizations) used in quantitative studies of international relations. It also provides users the ability to construct units of analysis ranging from monads (e.g., country-year), to dyads (e.g., country1-country2-year), to extra-dyadic observations called k-ads (e.g., country1-country2-year, . . . , -countryk-year)."

`peacesciencer` package for `R`
 	
`https://github.com/svmiller/peacesciencer`
	"`peacesciencer` is an R package including various functions and data sets to allow easier analyses in the field of quantitative peace science. The goal is to provide an R package that reasonably approximates what made EUGene so attractive to scholars working in the field of quantitative peace science in the early 2000s."

World Economics and Politics Dataverse (Graham et al., 2018; Graham and Tucker, 2019)	"The World Economics and Politics Dataverse [. . .] provides researchers the ability to download custom datasets with information about the political and economic characteristics of countries. The country-year version of the Dataverse includes over 1000 variables, covering more than 180 countries from 1816-2018. In 2019, we have added a dyad-year version of the resource which provides information about the relationship between countries – e.g., geographic distance, trade flows, treaties, and wars."
Correlates of War (COW) `https://correlatesofwar.org/`	A collection of widely used, harmonized datasets, including: state system membership; militarized interstate disputes (MIDs); intergovernmental organization membership; formal alliances; diplomatic exchange; trade; national material capabilities; religious adherence; territorial contiguity and territorial change.
Paul Hensel's International Relations Data Site `http://www.paulhensel.org/data.html`	"This site includes seven pages of links to on-line data resources for the serious international relations scholar [. . .] These pages are meant to include the most useful data sources on processes of international conflict and cooperation, as well as data covering international economic, environmental, political, and social data and data on similar topics for the United States [. . .] [T]hese pages attempt to provide important information about each data set to help the user determine which data set(s) would be most useful."

`WDI` package for `R`
 	
`https://github.com/vincentarelbundock/WDI`
	"The WDI package allows users to search and download data from over 40 datasets hosted by the World Bank, including the World Development Indicators ('WDI'), International Debt Statistics, Doing Business, Human Capital Index, and Sub-national Poverty indicators."

`countrycode` package for `R` Arel-Bundock, Enevoldsen and Yetman (2018)	"International organizations, statistical agencies, and research labs use different codes to represent countries [. . .] When researchers merge and analyze data from several sources, incompatible country codes can be a major source of frustration. The `countrycode` package for `R` alleviates this problem[. . .] [I]t allows bidirectional conversion between more than 30 country code schemes [. . .] [and] can convert codes into the names of countries in almost any spoken language."
`kountry` package for Stata Raciborski (2008)	"A Stata utility for merging cross-country data from multiple sources [. . .] can be used to translate one country-coding scheme into another, to recode country names into a 'standardized form', and to generate geographic-region variables."

NewGene (Bennett, Poast and Stam, 2019) is a stand-alone program that allows users to construct datasets by automatically merging together variables from various sources. It is a recent redesign of the EUGene software (Bennett and Stam, 2000), originally released in 1998. The software comes pre-loaded with a large number of datasets, and also allows users to import other datasets not already included. The NewGene program creates datasets which users must then transfer into a separate program for analysis. The `peacesciencer` package is an effort to streamline the process, allowing users to create these datasets directly in the `R` environment in a similar manner as they would in NewGene. As of this writing, the `peacesciencer` package does not have as expansive data coverage or functionality as does NewGene, but it is being regularly updated and expanded. The World Economics and Politics (WEP) Dataverse (Graham et al., 2018; Graham and Tucker, 2019) performs a similar function to NewGene but through a web-based interface, and with slightly less expansive coverage and functionality.

The next three entries in the table are aggregations of datasets which can assist researchers in their search for variables to use in their analyses. The Correlates of War Project hosts a number of widely used IR datasets. Paul Hensel's International Relations Data Site is an enormous directory of links to and information about datasets hosted on other websites. Vincent Arel-Bundock's `WDI` package is an `R` package which enables users to import and merge various World Bank datasets directly into `R` (like the `peacesciencer` package, but with coverage limited to the World Bank datasets). Finally, we point readers to the `countrycode` package in `R`, and the `kountry` package in Stata, as resources that can help reconcile country names and coding schemes from different data sources.

5.2 Costs and benefits of automation

The resources described above provide some obvious benefits to the applied IR researcher. "The great value of EUGene," according to one of its creators, "lies in its automation of several tedious, potentially error-prone, and repetitive parts of the research process" (Bennett, 2011). As a result, the researcher becomes less likely to make mistakes in the early stages of the data merging process which would propagate through the analyses. In addition, automating these front-end tasks can free up more of the researcher's limited time and energy to the more intellectually demanding aspects of the research process. The creators of the WEP Dataverse likewise note that "[b]y standardizing the process by which diverse data sources are cleaned and merged together, the Dataverse also reduces the opportunities for data management errors." They suggest some additional benefits as well:

> The WEP Dataverse serves the research community by making it easier for researchers to access information that is otherwise spread across dozens of distinct sources. The WEP Dataverse allows researchers to evaluate whether their findings are robust to alternative measures of key concepts; it also enables them to reproduce, challenge, and extend the findings of others, making social science more rigorous.[7]

Yet theses resources also carry some downside risk for the resulting quality of the research product. First, as Bennett (2011) acknowledges: "If users do not understand the underlying data set structure and the underlying data collection, this may lead to the misuse of data. A particular issue regarding never having to examine individual data sets may be that users never have to actually look at data sets' codebooks." Precisely because of the value that EUGene (and now NewGene and other programs) provide, it becomes easier for researchers to carelessly employ these tools without thinking through the decisions that go into the construction of their datasets. When using these programs, it is essential that researchers only outsource the menial parts of the process to the software program, and not outsource the thinking and deliberation which they would otherwise

[7] `https://ncgg.princeton.edu/wep/about.html`

undertake. We recommend that when using these programs, researchers should still closely examine the automated decisions regarding complicated cases (e.g. cases of state independence, secession, or (re-)unification)—and of course, read the codebooks of any variables they use.

A second concern arising from the automated construction of datasets for analysis is related to one of the benefits identified in the above quote from the WEP website. Consider a research question for which one of the key concepts has multiple plausible operationalizations. Returning to our collection of papers in Table 1, we see several cross-national studies whose outcome of interest is some measure of the representativeness of government. These include various measures from the Polity project (Marshall, Jaggers and Gurr, 2002), such as the "Polity2" index ranging from -10 to 10, the "democracy" index ranging form 0 to 10, or individual components of those indices, such as "constraints on the executive" or "political competition"; a measure of "checks" from the Database of Political Institutions (Cruz, Keefer and Scartascini, 2018); the human empowerment index from CIRI (Cingranelli, Richards and Clay, 2014); and measures of political competition from the V-Dem (Pemstein et al., 2018) and Polyarchy (Vanhanen, 2002) datasets.[8] The availability of a variety of plausible outcome measures to choose from opens up a wide range of "researcher degrees-of-freedom", and the software programs discussed above make it all the easier for researchers to fully exploit those degrees of freedom.

The WEP creators point out that their resource "allows researchers to evaluate whether their findings are robust to alternative measures of key concepts"; more cynically, it also enables researchers to more easily p-hack or fish for results that conform to their theory, and then develop ex-post justifications for why they chose the variables that they did. (With enough imperfectly correlated variables to choose from, the probability of getting a "significant" finding with at least one of them, by random chance, approaches one.) Unfortunately there is no clear mechanism available to prevent such behavior, beyond the researcher's own conscience (and an understanding that such behavior is scientific fraud). Perhaps as a systematic, long-term response, the discipline can move towards normalization of reporting results with a wider range of alternative measures of key variables, along with a shared appreciation of the fact that not all results will be robust to all possible measures—and that such a standard should not be necessary for publication. We will bracket these sorts of ethical and industrial considerations and simply note that these are issues that researchers should remain mindful of.

5.3 Choosing variables

With the various resources available to find measures of theoretical concepts and construct datasets for analysis, how should we go about selecting the variables we want to use? This is certainly a challenging aspect of the research process, and there is no universal or formulaic solution to apply. Our general procedural advice is to start from the pre-analysis plan. Before merging any variables into the dataset—and of course, before running any regressions—the researcher should work through the exercise of thoroughly assessing and comparing among the options available. When faced with multiple variables to select from, the researcher should consider the following questions:

- How are the different variables defined and coded, according to the codebooks or articles accompanying the datasets?
- How do other studies that employ each variable explain their decision to do so?
- How do the variables differ in their coverage (temporal, spatial, and degree of missingness)?

[8] Freedom House (`https://freedomhouse.org/countries/nations-transit/scores`) provides another measure of democracy which is used frequently in IR and comparative politics, though not by any of the papers appearing in Table 1.

- What is the relative importance of each of the above considerations (insofar as there are trade-offs between the various measures)? For instance, is it more important to use a measure that more closely represents the theoretical concept, or to generate results that are directly comparable to existing literature? Do you prefer a variable with less measurement error and more limited coverage, or vice-versa?

We suggest these questions as a starting point for the researcher when selecting among available measures, while recognizing that answers to these questions will come more easily in some cases than in others. For certain research topics, lively scholarly debates have emerged regarding the choices of variables and relative merits of different datasets. On measures of democracy, for instance, Casper and Tufis (2003) and Cheibub, Gandhi and Vreeland (2010) examine the extent to which different empirical results are robust to alternative democracy indices. Boese (2019) provides a thorough comparison of the measures themselves, concluding that "The measures developed by the V-Dem project outperform Polity2 and Freedom House Index (FHI) with respect to the underlying definition and measurement scale as well as the theoretical justification of the aggregation procedure." Questions around measurement of international conflict and military intervention have similarly spurred much debate; see for instance Fordham and Sarver (2001) and Pickering and Kisangani (2009), and Gibler, Miller and Little (2016), Palmer et al. (2020), and Gibler, Miller and Little (2020).

As alluded to above, the ever-increasing accessibility of alternative measures makes it increasingly easy for the researcher to run her analyses with all possible measures and compare results. This practice is unproblematic if done transparently: that is, if the researcher reports all such results (not only the significant ones) and does not claim to have developed hypotheses ex-ante that she in fact arrived at ex-post. In the event that the researcher does choose to test multiple alternative measures, however, the comparative predictions should be developed in a pre-analysis plan. How do the measures differ from one another, and how should we expect those differences to bear on the statistical results? Would a positive finding with one measure but not another be informative of the operative theoretical mechanism? Can the differences in results be attributed to differences in the accuracy or coverage of the various measures? Thinking through these questions before conducting the analyses will invariably produce more valuable scientific insights than simply rationalizing the results after observing them.

Finally, we note that the preceding discussion focused on cross-national datasets that are publicly available and easily incorporated into a standard country-year (or k-ad-year) dataset for analysis. In many applications there is no such dataset available (let alone multiple such datasets which the researcher can choose from). In this case, the researcher will have to be creative in seeking out measures from novel sources. For some examples from our selection of studies in Table 1: Crost, Felter and Johnston (2014) and Sexton (2016) draw their measures of conflict incidents directly from reports of the respective militaries involved; Dube and Naidu (2015) use measures of violence gathered from local newspapers and from oral reports from networks of Catholic priests; Brollo et al. (2013) extract measures of mayoral corruption from reports of audits undertaken by the federal government of Brazil. There are considerable risks and rewards to constructing a novel dataset, as discussed in depth in the chapter by Braithwaite (2021) in this text.

5.4 Variable transformations

Beyond the question of which data source to select a variable from, the researcher must also confront the question of whether and how to transform the variable for analysis. We touch on these issues briefly here, and refer readers to other chapters of this volume for a more thorough treatment.

A frequent question that arises when operationalizing a variable is the choice of whether or not to discretize the variable: that is, given an interval measure, or an ordinal measure with many values, should the researcher collapse (or "bin") the variable into a smaller number of ordinal

values? Returning to the example of democracy measures, the Polity2 index is a 21-point scale, taking on values from -10 to 10. The researcher may analyze the index as an interval variable, such that a given change is treated as equivalent at any point in the scale (moving from -8 to -5 is equivalent to moving from 3 to 6, and so on). Alternatively, the researcher may dichotomize the variable, selecting a cutpoint in the -10 to 10 range such that all values above the cutpoint are labeled "democracy" and all values below are labeled "autocracy", and analyze democracy as a binary measure. Or the researcher may trichotomize, yielding a three-level ordinal variable with values "democracy", "anocracy", and "autocracy" (or she may polychotomize with more than three values).

As a general concern, discretizing a variable could mean losing information, and thus increasing the variability of the estimates and diminishing statistical power. It may also produce results that are sensitive to the cutpoints used for separating the levels of the discrete variable. Alternatively, leaving a variable as an interval measure could mean treating unequal changes as equal—moving from -8 to -5 on the Polity2 index may be qualitatively different from moving from 3 to 6—and it could mean ignoring discontinuities or jumps that exist in the theoretical quantity of interest. For a more thorough examination of the choice between binary, ordinal, and continuous measures of democracy, see Cheibub, Gandhi and Vreeland (2010). For a more general discussion of discretizing variables for analysis, see Gelman and Park (2009). The question has also been of particular interest as it applies to subgroup analyses and interactive effects; see Hainmueller, Mummolo and Xu (2019).

6 Measurement

The next question we consider pertains to the nature and extent of measurement error in the variables used in the analysis. As discussed in the previous section, this is one important consideration that should factor into the decision over variable selection. Here we provide a more detailed discussion of the implications of different forms of measurement error.

6.1 Classical measurement error

To begin, consider the simplest and most innocuous case of "classical" or random measurement error. This is a situation in which a variable is measured with some degree of random noise which is uncorrelated with the variable itself (or with other variables of interest). The implications of this form of measurement error vary depending on what function the variable is serving in the analysis. For concreteness, consider a version of our research question on windfalls and governance, where the researcher specifies the following regression equation:

$$Democracy_i = Foreign\,Aid\,Receipts_i\,\beta + pcGDP_i\,\gamma + e_i$$

The measure of fiscal windfalls in this case is a government's foreign aid receipts. For simplicity, suppose that conditional on GDP per capita, foreign aid receipts are exogenous to democracy, such that controlling for GDP per capita is sufficient to identify the effect of foreign aid on democracy. The researcher wants to test a "resource curse" theory of foreign aid, predicting that foreign aid will cause recipient governments to become less democratic.

What will be the consequence of classical measurement error in each of the three variables in the analysis? First, if there is solely measurement error in the dependent variable, democracy, it can simply be treated as an additive component of the error term e_i. The resulting estimation of β will be unbiased, but less efficient (meaning there will be more uncertainty and a larger standard error).

If instead there is measurement error in the independent variable, foreign aid receipts, the impact will be to bias the estimate of β towards zero. Intuitively, we can think of this as arising

from the fact that values of "high" foreign aid receipts have been miscoded as "low", and vice-versa. Assuming that the relationship between "true" aid receipts and democracy is negative (so truly high aid receipts have a low average democracy score, and vice versa), the observed high aid values (some of which are truly low values) will be associated with higher average democracy values, relative to the true aid values; thus the slope of the relationship between observed aid and democracy will be flattened. In the limit, the slope goes to zero as the measurement of aid becomes arbitrarily noisy.

Finally, the impact of classical measurement error in the control variable, GDP per capita, will depend on the sign of its relationship between both the dependent and independent variables. If income is negatively correlated with aid receipts and positively correlated with democracy, then by the Frisch-Waugh Lovell (FWL) theorem, failing to adequately control for income will result in an over-estimation of the (negative) effect of aid on democracy. Put simply, we should expect the resulting estimate of β to be somewhere between the true value of β, and what we would estimate if we omitted GDP per capita as a control variable entirely.

6.2 "Non-classical" measurement error

What about "non-classical" measurement error—that is, measurement error that is correlated with some other variable of interest? As a general rule, this is a far more complicated problem to address than classical measurement error. In the simplest cases, the research can tell a story to explain how measurement error matters, and "sign the bias" that arises from it. For instance, in our aid and democracy example: if autocratic governments are systematically under-reporting their aid receipts, or donors are systematically under-reporting their aid disbursements to autocratic recipients, this would a form of measurement error that is correlated with the theoretical quantities of interest. The result in this case would be to bias the estimated effect of aid toward zero, providing a "conservative" estimate; if we are confident that the measurement error only operates in the direction specified, then we can infer that our estimated effect of aid on democracy is a lower bound (in absolute value terms). Conversely, if democracies under-reported their aid receipts, or if autocracies over-reported them, then the bias would go in the same direction as our estimated effect, and we should infer that the true effect of aid is smaller in magnitude than what was estimated.

It can be appealing to tell these kinds of stories regarding the direction of bias arising from measurement error in the simplest cases. However, it is much more difficult to generalize the same logic to a multivariate setup, or to more complicated empirical models—for instance, models with interaction terms, or with first-differenced variables, as we see in many of the studies in Table 1. Our general recommendation is to seek out variables with a minimal degree of measurement error, rather than trying to make the case that the measurement error is innocuous, or that it is working in the opposite direction of a hypothesized effect.

With that goal in mind, how can a researcher know the type and extent of measurement error in a given measure? Of course, by its very nature, measurement error is unobservable—it is already baked in to the measure under observation, and inseparable from the "true" underlying values. But beyond just speculation or intuition, there are steps the researcher can take to get a better understanding of the problem. First, the researcher should review any related methodological literature which may inform her decision, as well as any pertinent discussion in existing studies that employ the measures she is considering. Second, the researcher should closely examine the codebook for the accompanying dataset, and consider the process by which the data are reported, collected, and coded, and where errors may arise in the process. Do the actors involved in any step of the process have incentives to hide or misrepresent the truth? Are there capacity limitations that prevent accurate measurement, independent of any intentional misreporting or miscoding of the data? Are there explicitly stated coding decisions which the researcher disagrees with, or which are inappropriate for the research question at hand? As one example, Boese (2019) highlights the systematic measurement error in the very commonly used Polity2 score from the Polity dataset,

whereby cases of "interregnum" (that is, collapse of central authority) are coded as 0, which is also the score assigned to midpoint between full democracy (+10) and full autocracy (-10). It seems likely that researchers would often disagree with this coding decision in applications.

7 Missingness

We will next consider the question of missing data. This is a pervasive problem in cross-national time-series datasets, which can have major ramifications for the internal and external validity of the resulting analyses. Due to limitations of space, we will discuss these issues briefly here, and refer readers to the chapter by Barnum (2021) in this volume for a more thorough treatment.

As in the discussion of measurement error, we will first note that not all forms of missingness are equally concerning. Generally speaking, we can separate the problem into cases of "missingness at random" and "missingness not at random". Consider an observation i in the sample of analysis, and let $m_{x,i}$ be an indicator denoting that the value of variable x for observation i is missing ($m_{x,i} = 1$) or not missing ($m_{x,i} = 0$) from the dataset. Missingness at random refers to a situation in which $m_{x,i}$ is unrelated to any other variables. When data are missing at random, the only problem is a loss of efficiency; no bias is introduced, but the estimation will simply be less precise due to the decrease in sample size.

The case of non-randomly missing data is the more challenging one. We can further divide this case into two subcases: missingness as a function of "pre-treatment" covariates, and missingness as a function of "post-treatment" covariates. The former case means that $m_{x,i}$ may be related to some background characteristics of observation i, but is not influenced directly by the treatment variable (the main independent variable of interest), or by any other variables which are in turn influenced by the treatment variable. In a country-year dataset, this might characterize a situation in which a given country is missing entirely from the dataset, or in which country-years with smaller populations or lower development levels are more likely to have missing values of the variable of interest. This form of missingness will limit the external validity of the analysis, but generally will not undermine its internal validity; that is, the resulting estimates should be unbiased for the parameters of the population which the sample of analysis resembles, but will be of limited value in informing us about the types of countries which are outside of that scope.

Missingness on the basis of post-treatment variables is far more concerning, and can undermine both internal and external validity of the analyses. For a thorough treatment of the topic, we recommend Montgomery, Nyhan and Torres (2018). For intuition, consider how this issue would emerge in our running example of windfalls and governance. Suppose that fiscal windfalls cause incumbent leaders to become more corrupt and less responsive to the public interest. It is entirely plausible that, as a consequence, these leaders become less likely to publicly report economic data, or to share such data with international organizations (which would compile the data and disseminate them for use by researchers) (Hollyer, Rosendorff and Vreeland, 2018). If a researcher were to use these economic variables as control variables in her analysis, and did nothing to correct for the missing values, then country-years for which governance quality has deteriorated due to fiscal windfalls will be systematically more likely to be omitted from the analysis. Clearly this yields a biased estimate of the effect of windfalls on governance.

Recommendations to the problem of missing data are varied, and are unfortunately beyond the scope of this chapter. One popular (but by no means uncontroversial) approach, known as "multiple imputation", involves predicting the values of the missing observations and using those predicted values in the analysis. See Honaker and King (2010), Arel-Bundock and Pelc (2018), Pepinsky (2018), and Barnum (2021) for more discussion of the relative merits and drawbacks of multiple imputation.

8 Statistical power

The final topic we consider is the statistical power of an empirical analysis. A power analysis is a standard feature of a pre-analysis plan for an experimental study; as with the other topics addressed throughout this chapter, we suggest that conducting a power analysis can be similarly valuable for the applied IR researcher using observational data.

Statistical power, generally speaking, is the probability that a statistical test will reject the null hypothesis, given that the alternative hypothesis is in fact correct. The purpose of a power analysis is to answer the following question: Suppose that the true effect size of our treatment of interest is some value τ. Suppose that we will conduct a statistical test of size α (typically $\alpha = 0.05$, for a hypothesis test with 95% confidence). Given the variability of our outcome variable, the treatment assignment mechanism, the model specification, and the sample size, how likely are we to detect an effect that is distinguishable from zero? In other words, assuming our treatment does in fact have the effect that we hypothesize it to have, what is the probability that our research design will succeed in producing a statistically significant result?

An answer to this question is perhaps of the most obvious value to the experimental researcher, who has a greater degree of control over the various research design features listed above (most notably the treatment assignment mechanism and the sample size). For the observational researcher, however, this information is still valuable to know before conducting her analyses. First, as mentioned above, understanding the power of a given research design can help the researcher adjudicate between different operationalizations of the variables of interest. It can likewise inform her decisions regarding the appropriate scope of the analyses, and how many countries and years should be included in the sample. Further, knowing the power of a research design can be indicative of the scientific value of a null result (that is, a failure to reject the null): a null result of a well-powered test is more informative than a null result of an under-powered test. Knowing the power of the design, and perhaps revising the design to improve its power, can help ensure that the study will be of interest to the broader scientific community regardless of the particular results it generates.

The general procedure for conducting a power analysis is as follows. The research first sets parameters of the research design as described above (sample size, effect size, outcome variation, treatment assignment probabilities, and so on). She then simulates data according to these parameters, specifies the statistical model, and runs the analysis on the simulated data. This process can then be repeated across a range of parameter values of interest. Applying this procedure to observational IR data, the researcher might choose instead to use her real data for the righthand-side variables (treatment variable and covariates), while using simulated or "scrambled" data for her outcome variable. This would enable her to explore the impact of different decisions regarding model specification, variable operationalization, and sample selection, on the power of her design. A more detailed guide for conducting power analyses can be found in Gerber and Green (2012). A recent and more holistic approach to declaring and pre-analyzing research designs is discussed in Blair et al. (2019).

An important consideration in power analyses for conventional IR datasets—and of course, in the analyses themselves—is the existence of dependencies among observations. A naive approach to statistical inference with a dataset consisting of country-year observations would treat each observation as if it were contributing its own independent piece of information. However, there is generally good reason to believe that these observations are not independent, but rather are related within units over time, or across units within a period of time. Ignoring these dependencies will usually (though not always) result in an underestimation of the degree of uncertainty in the parameter estimates of interests (i.e. yielding standard errors that are too small), and thus an overstatement of the degree of confidence in the results (i.e. an increased risk of falsely rejecting a null hypothesis). The conventional approach to accounting for these dependencies is to use a Cluster-Robust Variance Estimator (CRVE), or more colloquially, "clustered standard errors"; with

country-year data, researchers often cluster standard errors by country, though we see no reason why researchers should not instead use a more robust approach of clustering by both country and year. For an overview of cluster-robust inference, see Angrist and Pischke (2008, ch.8) and Cameron and Miller (2015). For additional complications that arise in a dyadic setting, see Aronow, Samii and Assenova (2015). Incorporating cluster-robust standard errors into a power analysis will make evident that a dataset which appears on its face to have 5,000 observations may actually be better understood as having closer to 100 observations, for the purposes of statistical inference.

Working through the power analysis process as outlined above provides further benefits for the researcher beyond merely providing calculations of statistical power. As Humphreys, Sanchez de la Sierra and Van der Windt (2013) describe their experience with conducting analyses of simulated data and writing up a "mock report" of the results, the process

> forced us to take early action on the plethora of small decisions that can provide the latitude for fishing (intentional or not)... [I]t required absolute precise definitions of dependent variables and choices over dichotomizations... [I]t forced a selection of covariates... [and] forced a selection of subgroup analyses.

Incorporating these practices into the research process can prove valuable for the observational IR scholar as well as the experimental researcher.

9 Conclusion

This chapter has sought to provide general guidance on the steps of the research process in between theory development and statistical analysis for researchers using conventional IR data. We have discussed the decisions that must be made regarding the selection of a sample of analysis and the operationalization of variables, and the implications of these choices for internal validity, external validity, and statistical power. We suggest that applied IR scholars can benefit from adopting a range of practices used by experimental researchers, which help to develop a structured approach for thinking through all aspects of the research design before actually conducting the analyses. These practices become all the more valuable as the datasets and tools for quantitative analysis of international politics become more easily accessible. Confronting these challenging research design questions in advance will improve both the process and the product of our research.

References

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arel-Bundock, Vincent and Krzysztof J Pelc. 2018. "When can multiple imputation improve regression estimates?" *Political Analysis* 26(2):240–245.

Arel-Bundock, Vincent, Nils Enevoldsen and CJ Yetman. 2018. "countrycode: An R package to convert country names and country codes." *Journal of Open Source Software* 3(28):848.

Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

Aronow, Peter M, Cyrus Samii and Valentina A Assenova. 2015. "Cluster–robust variance estimation for dyadic data." *Political Analysis* pp. 564–577.

Balcazar, Carlos Felipe and Rafael Ch. 2021. "Do Tariff Revenues Generate A Resource Curse? Theory And Evidence From The First Wave Of Globalization." New York University, MIMEO.

Barnum, Miriam. 2021. Dealing with Missing and Incomplete Data. In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Bennett, D Scott. 2011. "Is EUGene a Collective Bad?" *Conflict management and peace science* 28(4):315–330.

Bennett, D Scott and Allan C Stam. 2000. "EUGene: A conceptual manual." *International interactions* 26(2):179–204.

Bennett, D Scott, Paul Poast and Allan C Stam. 2019. "NewGene: An Introduction for Users." *Journal of Conflict Resolution* 63(6):1579–1592.

Berset, Simon and Mark Schelker. 2020. "Fiscal windfall curse." *European Economic Review* 130:103592.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113(3):838–859.

Boese, Vanessa A. 2019. "How (not) to measure democracy." *International Area Studies Review* 22(2):95–127.

Braithwaite, Jess. 2021. Challenges and Payoffs of Building a Dataset from Scratch. In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti and Guido Tabellini. 2013. "The political resource curse." *American Economic Review* 103(5):1759–96.

Brückner, Markus, Antonio Ciccone and Andrea Tesei. 2012. "Oil price shocks, income, and democracy." *Review of Economics and Statistics* 94(2):389–399.

Bueno de Mesquita, Bruce and Alastair Smith. 2013. "Aid: Blame it all on "easy money"." *Journal of Conflict Resolution* 57(3):524–537.

Butler, Chris. 2021. What kind of analysis is appropriate for my data? Should we just use OLS for everything? In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Cameron, A Colin and Douglas L Miller. 2015. "A practitioner's guide to cluster-robust inference." *Journal of human resources* 50(2):317–372.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign aid, human rights, and democracy promotion: Evidence from a natural experiment." *American Journal of Political Science* 61(3):671–683.

Caselli, Francesco and Andrea Tesei. 2016. "Resource windfalls, political regimes, and political stability." *Review of Economics and Statistics* 98(3):573–590.

Casper, Gretchen and Claudiu Tufis. 2003. "Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated data sets." *Political Analysis* pp. 196–203.

Cheibub, José Antonio, Jennifer Gandhi and James Raymond Vreeland. 2010. "Democracy and dictatorship revisited." *Public choice* 143(1):67–101.

Chiozza, Giacomo. 2021. Regression Analysis. In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Christensen, Garret and Edward Miguel. 2020. Transparency and Reproducability: Potential Solutions. In *The Production of Knowledge: Enhancing Progress in Social Science*, ed. James Mahoney Colin Elman, John Gerring. Cambridge University Press pp. 165–196.

Cingranelli, David L., David L. Richards and K. Chad Clay. 2014. "The Ciri Human Rights Dataset.". `http://www.humanrightsdata.com`.

Correlates of War Project. 2017. "State System Membership List, v2016.". `http://correlatesofwar.org`.

Crost, Benjamin, Joseph Felter and Patrick Johnston. 2014. "Aid under fire: Development projects and civil conflict." *American Economic Review* 104(6):1833–56.

Cruz, Cesi, Philip Keefer and Carlos Scartascini. 2018. "Database of Political Institutions 2017 (DPI2017)." Inter-American Development Bank. Numbers for Development. `https://mydata.iadb.org/Reform-Modernization-of-the-State/Database-of-Political-Institutions-2017/938i-s2bw`.

Djankov, Simeon, Jose G Montalvo and Marta Reynal-Querol. 2008. "The curse of aid." *Journal of economic growth* 13(3):169–194.

Dube, Oeindrila and Juan F Vargas. 2013. "Commodity price shocks and civil conflict: Evidence from Colombia." *The review of economic studies* 80(4):1384–1421.

Dube, Oeindrila and Suresh Naidu. 2015. "Bases, bullets, and ballots: The effect of US military aid on political conflict in Colombia." *The Journal of Politics* 77(1):249–267.

Fordham, Benjamin O and Christopher C Sarver. 2001. "Militarized interstate disputes and United States uses of force." *International Studies Quarterly* 45(3):455–466.

Gelman, Andrew. 2013. "Preregistration of studies and mock reports." *Political Analysis* 21(1):40–41.

Gelman, Andrew and David K Park. 2009. "Splitting a predictor at the upper quarter or third and the lower quarter or third." *The American Statistician* 63(1):1–8.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation.* WW Norton.

Gibler, Douglas M, Steven V Miller and Erin K Little. 2016. "An analysis of the militarized interstate dispute (MID) dataset, 1816–2001." *International Studies Quarterly* 60(4):719–730.

Gibler, Douglas M, Steven V Miller and Erin K Little. 2020. "The Importance of Correct Measurement: A Response to Palmer, et al." *International Studies Quarterly* 64(2):476–479.

Gill, Jeff. 2001. "Whose variance is it anyway? Interpreting empirical models with state-level data." *State Politics & Policy Quarterly* 1(3):318–338.

Goldberg, Ellis, Erik Wibbels and Eric Mvukiyehe. 2008. "Lessons from strange cases: Democracy, development, and the resource curse in the US states." *Comparative Political Studies* 41(4-5):477–514.

Gonzalez, Elsy and Paul Poast. 2021. What kind of data is appropriate for my question? Choosing a Unit of Analysis. In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Graham, Benjamin A. T., Raymond Hicks, Helen Milner and Lori D. Bougher. 2018. "World Economics and Politics Dataverse.". `https://ncgg.princeton.edu/wep/dataverse.html`.

Graham, Benjamin AT and Jacob R Tucker. 2019. "The international political economy data resource." *The Review of International Organizations* 14(1):149–161.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27(2):163–192.

Hollyer, James R, B Peter Rosendorff and James Raymond Vreeland. 2018. *Transparency, Democracy, and Autocracy: Economic Transparency and Political (In) Stability.* Cambridge University Press.

Honaker, James and Gary King. 2010. "What to do about missing values in time-series cross-section data." *American journal of political science* 54(2):561–581.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter Van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* pp. 1–20.

Janzen, Sarah A and Jeffrey D Michler. 2021. "Ulysses' pact or Ulysses' raft: Using pre-analysis plans in experimental and nonexperimental research." *Applied Economic Perspectives and Policy* .

Lemke, Douglas and William Reed. 2001. "The relevance of politically relevant dyads." *Journal of Conflict Resolution* 45(1):126–144.

Marshall, Monty G., Keith Jaggers and Ted Robert Gurr. 2002. "Polity IV project: Political regime characteristics and transitions, 1800-2002.".

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on post-treatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Nielsen, Richard A, Michael G Findley, Zachary S Davis, Tara Candland and Daniel L Nielson. 2011. "Foreign aid shocks as a cause of violent armed conflict." *American Journal of Political Science* 55(2):219–232.

Nunn, Nathan and Nancy Qian. 2014. "US food aid and civil conflict." *American Economic Review* 104(6):1630–66.

Palmer, Glenn, Vito D'Orazio, Michael R Kenwick and Roseanne W McManus. 2020. "Updating the militarized interstate dispute data: A response to Gibler, Miller, and Little." *International Studies Quarterly* 64(2):469–475.

Pemstein, Daniel, Kyle L Marquardt, Eitan Tzelgov, Yi-ting Wang, Joshua Krusell and Farhad Miri. 2018. "The V-dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data." *V-Dem Working Paper* 21.

Pepinsky, Thomas B. 2018. "A note on listwise deletion versus multiple imputation." *Political Analysis* 26(4):480–488.

Pickering, Jeffrey and Emizet F Kisangani. 2009. "The International Military Intervention dataset: An updated resource for conflict scholars." *Journal of peace research* 46(4):589–599.

Quackenbush, Stephen L. 2006. "Identifying opportunity for conflict: Politically active dyads." *Conflict Management and Peace Science* 23(1):37–51.

Raciborski, Rafal. 2008. "kountry: A Stata utility for merging cross-country data from multiple sources." *The Stata Journal* 8(3):390–400.

Ramsay, Kristopher W. 2011. "Revisiting the resource curse: Natural disasters, the price of oil, and democracy." *International Organization* pp. 507–529.

Ritter, Emily Hencken. 2021. Using Theory to Choose a Research Strategy (a.k.a. Matching Empirics with Theory or EITM). In *Handbook of Research Methods in International Relations*, ed. R. Joseph Huddleston, Tom Jamieson and Patrick James. Edward Elgar Publishing pp. XX–XX.

Sexton, Renard. 2016. "Aid as a tool against insurgency: Evidence from contested and controlled territory in Afghanistan." *American Political Science Review* 110(4):731–749.

Vanhanen, Tatu. 2002. "Polyarchy dataset.". `https://www.prio.org/Data/Governance/Vanhanens-index-of-democracy/`.

Western, Bruce and Simon Jackman. 1994. "Bayesian inference for comparative research." *American Political Science Review* pp. 412–423.