How to Code[*]

Thomas B. Pepinsky

Assistant Professor
Department of Political Science
University of Colorado-Boulder
pepinsky@colorado.edu

First (Preliminary) Draft: May 7, 2007

[*] Prepared for the 2007 Annual Meeting of the American Political Science Association, Chicago, IL. Thanks to the participants at the January 2006 Institute for Qualitative Research Methods for inspiring me to think about these issues.

How to Code

1. Introduction

Recent literature on social science research methodology focuses on issues of causal inference and hypothesis testing given existing data (Brady and Collier 2005; King et al. 1994). Another literature has focused on the antecedent question of concepts in social science research, along with the related questions of conceptual validity and conceptual stretching (Adcock and Collier 2001; Goertz 2005; Sartori 1970). The step that lies between the formulation of social science concepts and strategies of causal inference is coding. Despite its obvious importance, the methodological literature has yet to identify a standard process for turning concepts into variable that analysts can study using qualitative and/or quantitative methodologies. Yet clearly, both qualitative and quantitative researchers make implicit coding decisions in nearly all positivist (and much non-positivist) social science research. All too frequently, coding practices are opaque, ad hoc, and only weakly linked to concept or theory. This presents an obvious barrier to knowledge cumulation in the social sciences.

This paper proposes a "Coding Standard" to remedy this problem. The Coding Standard is a unified framework that unites the enterprise of coding conceptual variables for qualitative and quantitative researchers alike. There are four vital components of coding: theory, clarity, generality, and replicability. Theory is the conceptual basis of any coding scheme, and stipulates the possible values a variable can take and why (i.e. what is a democracy and what are its components?). Clarity refers to the existence of straightforward operational definitions of a variable's values, grounded in observable empirical facts (i.e., what phenomena must an analyst observe to identify the existence of "political protest"?). Generality refers to the ability of

coding schemes to travel outside of their original empirical domains, and if not, the provision of clear theoretical reasons why not (i.e., can one code government types from the "Varieties of Capitalism" literature in Latin America?). Replicability refers to the ability of a naïve researcher to recreate a coding decision or a dataset and achieve the same findings (if a new researcher attempted to recreate the Polity dataset, would she create an identical dataset?).

The paper proceeds in four sections. A brief review situates the problem of coding within the literature social science concepts and causal inference. The following section introduces the Coding Standard (hereafter, CS), systematically proceeding through the CS's four components and in several instances offering canonical examples of their successful application in real research. After introducing the CS, the subsequent section adjudicates thirty examples of empirical social science research that either implicitly or explicitly involve coding variables and scope conditions. Examples are drawn randomly from a sample of comparative politics and international relations journals between 2000 and 2006. The surprising finding regards small-n research: in sharp contrast to what practitioners often claim are the comparative strengths of small-n research, qualitative projects are often the most significant violators of best practices as identified by the CS. The final section concludes the paper with a discussion of its implications. Analysis of recent scholarship in comparative politics and international relations shows that weak coding practices are a common problem in both qualitative and quantitative work, suggesting that wider attention to the rigors and pitfalls of coding is a crucial step in building progressive research programs in social science. An unknown number of causal inferences in the social sciences are incorrect not because of problems of the procedure of inference, nor because of the muddiness of concepts, but simply because we have no way to adjudicate what we are studying.

2. Background: Concepts, Validation, and Inference

The proper techniques of empirical research have long been a subject of debate among social scientists. Disputes frequently arise between scholars who gather quantitative or qualitative evidence, or among scholars whose approaches are might be termed rationalist, behavioralist, or interpretivist. Within mainstream political science, though, there has been a remarkable (if unstated) meeting of minds among various schools (Hopf 2007 forthcoming). Whether or not researchers adhere to positivist methods of inference or endeavor to uncover contextual evidence of social processes, the vast majority of political scientists agree that the ultimate goal—however unattainable—of the empirical political scientist is to make true claims about the social world. This is a form of *ontological* positivism (as opposed to the epistemological positivism of a rational choice theorist): true facts exist, can be uncovered through careful analysis, and evidence that contradicts an existing theory means that the theory requires either rejection or refinement. This makes the task of parsing the social world into analyzable bits a fundamental component of social research. It is an endeavor shared by a participant-observer attempting to understand the discursive construction of mass political protest in an emerging democracy and an econometrician probing the covariation of short- and long-term interest rates alike.

In conceptualizing the process of coding, applied theorists have customarily focused on social science concepts, while methodologists have addressed on validity, reliability, and inference. One can envision this process as having an *ex ante* and an *ex post* component. Conceptual theorizing operates *ex ante*, as the empirical social scientist chooses the theoretical constructs which inform his analysis of social facts. Validation and inference happen *ex post*, after social facts have been assigned particular values of variables within a conceptual

framework. Despite the sub-disciplinary divide between the two enterprises, the process of coding *necessarily* includes both conceptual theorizing and validation.

2.1. Concepts

Concepts in social science are based on theory. Sartori (1970) was an early proponent of critical reflection on conceptual formation in political science, particularly comparative politics, where the proliferation of political systems across the world following the Second World War led to existing theoretical constructs being applied to empirical situations of little direct relevance. It made little sense to speak of, say, bureaucratic capacity—as theorized in the Western European and Northern Atlantic context—in many developing countries. Sartori urged scholars of comparative politics to refrain from stretching their concepts beyond their proper domains, to ensure that comparisons made between supposedly like units were actually useful comparisons. Using Lijphart's (1969) theory of consociational democracy, developed for the case of the Netherlands, to study politics in Switzerland, Malaysia, or Nigeria makes sense only if differences between these countries political systems truly can be ignored. Doing so requires a theory of what consociational democracy is, whose requirements can be ascertained in empirical domains aside from the inspirational or foundational case—in this case, the Netherlands.

Subsequent analysis of concept formation has refined Sartori's approach. Collier and Mahon (1993) argue that Sartori's view of concept formation and categorization is overly restrictive, necessitating categorical splitting when cases are not taxonomically identical. They argue that some concepts are better envisioned as having essential characteristics that do not necessarily overlap: a case may have four of five essential characteristics, but any four are sufficient for inclusion in a category. Goertz (2005) offers a more rigorous approach to concept formation, ranging from algebraic to Boolean systems. Variables may be scored according a mix

of necessary and/or sufficient conditions (elections, following Alvarez et al. 1996, are necessary but not sufficient for democracy), or as algebraic functions of existing data (dyadic democracy scores in most democratic peace research are the minimum of the two countries' levels of democracy). Both Collier and Mahon and Goertz stress the importance of theory in concept formation. Presumably, only theory tells researchers what the concepts are to be analyzed, and how to construct categories for analysis.

Debates over concept formation in political science are most often seen in research on regime types (Alvarez et al. 1996; Bollen and Jackman 1989; Collier and Adcock 1999; Collier and Levitsky 1997; Dahl 1971). Debates here focus specifically on the issue of concepts: what is a democracy, how do regime types vary, and how does one measure them? A host of coding schemes have subsequently attempted to operationalize these concepts into real variables (see Munck and Verkuilen 2002 for a review). The conceptual foundations of coding choices, though, are just as important for any type of qualitative or quantitative research in the social sciences.

2.2. Inference and Validation

Once analysts have coded variables, they use them to make inferences about the world. Debates on inference in social science remain prominent, with various methodologists staunchly defending case study narratives that shed light onto causal processes, others insisting upon quantitative approaches, and a more recent wave advocating multi-method approaches that bridge qualitative and quantitative methods (on these debates, see Brady and Collier 2005; George and Bennett 2005; Gerring 2007a; King et al. 1994; Mahoney and Rueschemeyer 2003; Sekhon 2004). Regardless of method, though, coding schemes that correctly map the social world onto a set of variables are instrumental for causal inference. Methodologists usually treat

poorly constructed variables as instances of measurement error. Quantitative methodologists have proposed a number of solutions for errors-in-variables regression models, including instrumental variables approaches (Fuller 1987) and structural equation models (Bollen 1989). Qualitative scholars have little to guide them through the pitfalls of miscoded variables. I return to this point below in assessing the consequences of unsatisfactory coding practices.

Validation—assessing the degree to which a variable measures the concept that it is intended to measure—is another *ex post* exercise given a newly coded variable. It has a long pedigree in psychometric research, but for political scientists, the most accessible introduction is Adcock and Collier (2001). It is clearly related to the process of coding, and indeed, the authors present a schematic of the coding process (Figure 1, pg. 531) that starts with general concepts and proceeds through operationalization of variables to validation. Operationalization is their gloss of what I have termed coding. For them, coding is assumed to be a straightforward, almost technical exercise.

But unstated in Adcock and Collier is the process through which good or bad measurements arise. Following their terminology, assume the existence of a perfect "background concept" as well as a perfect "systematized concept." Also assume that valid indicators are known theoretically. How does this turn into measurement—in other words, how does coding take place? How can social scientists be certain that coding choices are systematic, clear, and as error-free as possible? Adcock and Collier's methods of validation are entirely *ex post*, through the theoretical stipulation of sufficient information and nuance in the concept's indicators (content validation),[1] investigating the covariation between the new variable and existing variables that measure the same concept (convergent validation), and/or investigating

[1] Note the tension here. If the systematized concept is considered valid theoretically *ex ante*, then there should never be a case where *ex post* it is found to be invalid on theoretical grounds.

the covariation between the new variable and another variable measuring a different concept but whose relationship with the new variable is known (construct validation). None of these methods address the prior task of creating a coding scheme to measure the concept, which if done correctly, should render these *ex post* tests of validity superfluous. A valid measure, for instance, may well not correlate with existing indicators of the same concept for the simple reason that these existing indicators are poorly constructed. It may not follow existing correlations with a dependent variable because the existing relationship—only uncovered through a poorly-measured variable—is spurious, and dependent on a poorly constructed variable.

2.3. Consequences

I alluded previously to the customary approach of treating poorly constructed variables as variables measured with error. Familiar to most quantitative political scientists is the dictum that measurement error leads to biased coefficients in multiple regression. For this reason, poorly constructed variables whose values do not correspond to true social facts will accordingly yield biased inferences in any quantitative application. The stakes for qualitative research are even higher. Because qualitative scholars almost always suffer from negative degrees of freedom, inference takes place through careful tracing of causal logics (George and Bennett 2005) or through carefully chosen comparative cases (King et al. 1994; Przeworski and Teune 1970). In such settings, one improper variable coding—a dependent variable, an independent variable of theoretical interest, and even a confounding variable controlled for by case selection or the variable which codes the scope condition of the argument—renders inference impossible. Yet paradoxically, given the disproportionately serious consequences of measurement error for

qualitative research, the vast majority of methodological literature in political science that addresses errors in variable coding is *quantitative* in nature.

The consequences of poor coding may be more serious, however, than the simple measurement error model suggests. Statistical corrections for classical measurement error presume, among other technical assumptions about the distribution of measurement error conditional on true values, that there is a true concept in the social world that is being captured, imperfectly, by the observed variable. This is an assumption that cannot be tested or defended based on any statistical test—it relies solely on the rigor and parsimony of the theoretical concept. It is difficult, for example, to view Alvarez et al.'s (1996) coding of democracy as Bollen and Jackman's (1989) coding of liberal democracy "with error." The two make different theoretical claims about what democracy is, and while their measures are correlated, it is not the case that both claim to be measuring the same theoretical construct. So while coding requires technical precision to minimize measurement error, it also requires theoretical precision to ensure content validity. Inferences made based on one's data about "the effect of democracy, measured as X, on Y" will only be as valid as the (perhaps imperfectly measured) X's theoretical relation to democracy.

How extensive of a problem does coding present for social science research? Coding disputes inspire prominent articles in the most widely read and methodologically sophisticated journals in political science. An illustrative example is a forthcoming article by Cusack et al. (2007 forthcoming) in the *American Political Science Review*. The article directly challenges a well-known study by Boix (1999), also published in the *American Political Science Review*, of the origins of electoral systems in modern Europe. These studies are useful because the number of cases is small enough for straightforward qualitative investigation, but large enough for the

authors to have performed quantitative tests of their arguments. The debate between is both theoretical and empirical: Cusack et al. challenge both the logic of Boix's account and, in their words, "the interpretation of the historical record, and the cross-national evidence." They level specific charges:

> We believe there are issues with at least 6 of the 22 cases that form the main focus of Boix's empirical analysis. *First*, three cases are completely missing appropriate data prior to the introduction of PR: Finland, Greece, and Luxembourg. The specific issues in each case are discussed in Appendix B, but we think it is clearly inappropriate to include any of these cases. *Second*, there are three countries, Iceland, the Netherlands and Sweden, where the timing for the threat measure needs to be altered. In the case of the Netherlands, for example, Boix notes that he uses the year 1919, but there was no election in that year. Rather, there had been an election in 1918 where both PR and universal male suffrage were introduced simultaneously…
>
> The second measurement issue is the use of fragmentation on the right – defined as one divided by the effective number of right parties – as a measure of the absence of single right party dominance. A case where there is no dominant right party can have the same effective number of parties as a case where there is a clearly dominant right party (see section on the Boix model for a historical example). We therefore use a direct measure of single party dominance, namely the percentage lead of the largest right party over the next largest…
>
> Finally, it needs to be noted that in calculating the threat measures for the alternative years, the data we recovered from the original sources in some instances differed from Boix.

Cusack et al. thus argue that Boix has (1) coded cases that should fall outside of the empirical purview of the argument, (2) improperly measured the fragmentation of the right, and (3) simply coded several observations incorrectly. Note here that Cusack et al.'s criticisms of Boix's coding choices include both simple errors in coding (data from original sources differ) and more serious conceptual disagreements about the nature of the theoretical concept being measured (proper measure of single-party dominance differs) and even the empirical domains in which the argument should be tested (Finland, Greece, and Luxembourg; or not). This entails that, if Cusack et al. are correct, then one cannot

simply approach Boix's data with an errors-in-variables model. Cusack et al. adjust Boix's codings according to their own approach, and find that Boix's results no longer hold.

Remaining agnostic about which coding scheme is correct, it is clear that coding decisions have profound impacts on the knowledge that these authors have produced. If Cusack et al. are correct, then the innovation of Boix's article—the quantitative test of a formalization of Rokkan (1970) is invalid. Absent its empirical findings, it is unlikely that the article would have made the impact that it did. If Boix is correct, by contrast, than a major criticism of Cusack et al.—their reanalysis of the empirical data from Boix's earlier work—falls flat. Without Cusack et al.'s empirical critique, their theoretical critique is far less compelling.

Social scientific knowledge certainly progresses through such refinements and critiques. But these problems could be avoided if both sets of authors had followed a clearer, theoretically grounded, and easily replicable coding process. Had this been the case, the coding of cases could have been judged *ex ante*, and works then could have been evaluated on theoretical and empirical support for the arguments rather than largely on the believability of the procedure through which variables were constructed, for which no systematic metric exists.

3. The Coding Standard

Despite a proliferation of research on *ex ante* conceptual development and *ex post* inference and validation, a principled approach to coding in the social sciences remains missing. Coding is not merely a technical exercise, and authors continue to create variables and datasets

using opaque methodologies. The CS provides a metric by which to measure any variable's coding, and a guideline for researchers embarking on new coding projects.

3.1. Theory

The first component of any coding exercise is theory. Theory tells researchers the two most fundamental facts that are integral for coding, (1) what the concept being defined is, and (2) what its possible values are. These two facts determine not only what the social facts being observed are (see 3.2 below on operationalization and clarity), but also the structure of the variables: continuous, nominal, ordinal, or binary. Rigorous theoretical articulations of social science concepts are therefore instrumental for coding variables (Collier and Mahon 1993; Goertz 2005). In many social science applications, theoretical backgrounds for variables are well-known and well-accepted. Gross domestic product per year is the sum of consumption, investment, government expenditure, and net exports within a country during a given year. District magnitude refers to the total number of representatives that electors within a district vote into office during a election. For other concepts the relationship between theory and quantification remains far more contentious. The CS nevertheless demands that researchers develop clear theoretical articulations of their variables.

Methodologists usually emphasize the importance of theory for coding choices in the context of large-n datasets. But theory is just as important for qualitative research. Qualitative researchers should not believe that their deep knowledge of specific contexts makes them immune from having clear and well-defined theories of their variables, or that process-tracing does not involve implicit coding decisions. Sartori (1970: 1038) made a claim in his seminal piece on concepts that is no less relevant today as it was almost forty years ago: "concept formation stands prior to quantification." "Quantification" here means quite literally the creation

of *quanta*, or discrete units, of information. There is no divide between quantitative and qualitative coding in this regard, and his advice about the importance of theory is just as vital for qualitative scholars engaged in small-n research. Imagine, for example, a hypothetical study of the political ideologies that legitimize anti-regime protest in modern Zimbabwe. One might approach this problem by analyzing opposition texts, interviewing political activists, and engaging in participant observation. But theories of political ideology, anti-regime protest, and political legitimacy are instrumental for making sense of any findings. They are just as instrumental for guiding researchers to make good choices about what political ideology is (a coherent statement of a political platform, a positive articulation of an alternative, or a negative resistance to the hegemon?), what anti-regime protest is (open mass gatherings or subaltern resistance?), how legitimacy can vary (continuously or discretely, along one dimension or many?), among others. Many qualitative researchers ignore that in constructing their arguments, they have implicitly coded a set of variables. But forthright theoretical statements remain vital.

What does the importance of theory mean for best practices in coding? The first step in any coding project must be a clear formulation of the theoretical concept being measured. What is a war, a protest, a democracy, a capital market transaction, a parliament, an economic reform, a market economy, or an act of civil violence? Moreover, what are the values that these variables can take? Are wars discrete binary variables (at war or not), are all market economies coordinated or liberal, and what is the universe of types of civil violence that exist? This is particularly acute for new coding schemes that seek to draw together disparate data sources into new variables. Oftentimes, researchers approach data collection for new variables inductively, but without a theory to guide their enterprise, it will be impossible to ascertain the social realities

that variable attempt to capture. We can thus restate Sartori's dictum in starker terms: *No theory, no coding*.

As in Adcock and Collier (2001), there is clear room for an iterative process of definitional refinement. Scholars seeking to code Islamist political parties around the world based on the experiences of the Middle East will undoubtedly find new meanings for political Islam in South Asia and sub-Saharan Africa. Yet this dialogue between new cases and theory does not obviate the need for a final theoretical statement on the definition of an Islamist political party before making authoritative coding choices. The risk, otherwise, is definitional drift, with social facts not consistent with coding a political party as Islamist in Jordan used as criteria for coding a political party as Islamist in Nigeria.

3.2. Clarity

Clarity refers to the existence of clear statements about the social facts in the world that lead researchers to make coding decisions. A coder must be able to make statements of the following form: "when I observe fact A in the world, I code variable X as having value U." Or alternatively, for narrative approaches, "when I observe fact B in the world, I conclude that I have observed an instance of Y." The importance of theory is obvious for allowing coders to make such statements, but it is often at this step of variable operationalization that highly developed theoretical concepts become empirically vague. Perhaps the most well-known example of such issues is O'Donnell's (1973) work on bureaucratic authoritarianism, which developed a highly refined theoretical account of the rise and decline of such regimes based on the experiences of Southern Cone dictatorships in the 1960s and 1970s. Yet when moving from theory to case, finding the essential characteristics of bureaucratic authoritarianism that set it apart from other types of authoritarian rule (with names such as "praetorianism" and "neo-

feudalism," each just as poorly operationalized as bureaucratic authoritarianism) becomes entirely opaque. Without assertive statements about the social facts that lead one to classify a regime as bureaucratic authoritarian, it is simply impossible to ascertain whether or not bureaucratic authoritarianism as envisioned in the Argentine case truly applies to regimes in Brazil, Chile, or Uruguay, or further afield in the Middle East, Africa, and Asia.

Clarity places a tall order on the coder. In practice, many of the minutiae of coding decisions are arbitrary, and others difficult to justify on *a priori* grounds. But these are precisely those coding decisions which must be made as transparent as possible. Consider the daunting task of coding all instances of political violence within one city during one year. There will clearly be marginal cases of violence versus simple protest, and others where political violence overlaps with family, criminal, or other forms of violence. Coders engaged in the formation of a full dataset of acts of political violence will necessarily make decisions about which types of political action, or acts of violence, count as political violence. How should coders choose these coding rules? The answer, again, is theory. Rather than employing arbitrary coding rules, or developing coding rules as data becomes available, coders must remember that *clarity comes from theory, not from data*. If coders are unable articulate clear coding rules, this is a signal that their theoretical concepts are underdeveloped. At this point, the CS demands that researchers return to their theory before continuing with coding.

Returning again to the case of coding regime types, a paradigmatic example of clarity in coding is Alvarez et al. (1996), who present a set of transparent coding rules that allow them to map social facts about political systems onto a binary variable of regime type as democratic or not. Their coding is also notable because of their theoretical precision: they state their case of a theoretical conception of democracy as a binary variable, and then derive their coding rules

based on this theory. Instructive are several statements that contrast their approach with others. "Should we stick the cases which cannot be unambiguously classified, given our rules, into an 'intermediate' category, half way between democracy and dictatorship? This view strikes us as ludicrous. If we cannot classify some cases given our rules, all this means is that either we have bad rules or we have insufficient information to apply them" (21-22). They also contrast their approach with other approaches, such as Polity II (Gurr 1990), for whom clear definitions of key statements are impossible to reproduce. Whether or not one agrees with their theoretical justification for coding variables as they have, their rules are clear and transparent.

And again, qualitative researchers should not believe that their context-specific knowledge obviates their need for clear coding rules. In the hypothetical study of ideology, legitimacy, and protest in Zimbabwe, the researcher may have developed a nuanced understanding of Zimbabwean opposition politics, but clear coding rules still remain vital for allowing researchers to assess the empirical strength of her findings. A researcher unable to state clearly the coding rules that allow her to score variables in particular ways should be wary that her coding choices are either arbitrary, or reflect her own preferences for particular findings and inferences. The researcher should have theoretical reasons to believe, say, that legitimacy is a binary variable—it exists or it does not. She must then be able to articulate coding rules that allow her to classify situations, individuals, or whatever as either possessing or not possessing this characteristic. An example statement of this type would be "I code protests as being legitimate when I observe fact Y." Deep contextual knowledge of legitimacy in Zimbabwe may lead authors to denote this "fact Y" as a combination of several observable facts, perhaps none of them individually necessary or sufficient. Yet clarity remains nevertheless instrumental for the qualitative researcher wishing to establish empirical findings.

3.3. Generality

Generality requires a statement and a measure of the empirical scope of a coding scheme. The concept of generality thus moves from clear statements of variables, their values, and coding criteria to the scope of observations, cases, and evidence. In specifying the generality of one's coding scheme, one defines both the population of cases and relates this population to the sample of cases being analyzed. Doing so requires that coders envision the full range of possible cases for each variable, and either (1) exhaustively categorize all instances of the variable within that scope or (2) characterize the sampling strategy from which cases are drawn from the population. The generality of a coding scheme, like the theory and clarity, comes from theory. The scope may, of course, be narrow, but coders must nevertheless state and theoretically justify such limitations of the empirical scope of the coding scheme. In other words, the logical opposite of generality—specificity—is equally important for coding.

In the large-n context, methodologists have long noted the importance of knowing the relationship between observations and the population from which they have been drawn. Standard statistical techniques draw inferences based on the assumption of random selection of cases from the population. Methodologists have also developed a number of techniques for drawing inferences from non-random samples of data, including selection models (Heckman 1979) and models to estimate equations with truncated variables (Tobin 1958). Regardless of the method employed, a necessary assumption is that the sample selection criterion is either full, random, or at the very least predictable. Even imputation algorithms for incomplete data (Little and Rubin 1987; Schafer 1997) presume that the existing data matrix includes all observations in the sample, even if it contains missing data for some variables for individual observations. There

is no statistical solution for inference about data for which the relation of the sample to the population is unknown.

Qualitative social science methodologists have recently turned their sights upon understanding how individual cases fit into the larger population of cases from which researchers draw causal inferences (Fearon and Laitin 2004; Gerring 2004, 2007b). This is a welcome development, but their arguments can be pushed further. Qualitative researchers often select a small number of cases or instances of a particular phenomenon—say, communist political movements in Southeast Asia, or the pace of economic reforms in democratic Argentina. Inferences made on such studies depend critically on an understanding of how the cases relate to the empirical scope of the argument. Studying Southeast Asian communist movements, researchers will draw incorrect conclusions if the cases studied are only the violent or popular ones. If the researcher wishes to make inferences based only on these highly visible cases, this must be stated and defended. Studying economic reforms in Argentina requires a statement not only of what constitutes an economic reform and what values this variable can take (see 3.1 and 3.2 above), but also the empirical scope of cases of which Argentine reforms in the modern democratic era are an instance. Also, when does this era begin, and how long does it persist—what is the coding rule for the empirical scope variable? The researcher may believe that Argentina's frequent economic crises since 1980 make the country unique among emerging market economies. If this is the case, the researcher should justify this theoretically. If not, researchers should treat all inferences based on the Argentine case as tentative given the larger population of similar countries, from Chile to South Korea.

As best practice, then, before coding and as an integral component of theorizing and conceptual refinement, coder should *state the universe and the sampling rule, and justify each.*

Why has the coder chosen these cases, and what should analysts then conclude about the inferences drawn from these cases? Studies of regime types such as Polity IV (Marshall and Jaggers 2000) often truncate their data based on population, with their project including no countries with 2002 populations of less than 500,000 people. Such forthright disclosures of generality are crucial for determining the inferences researchers should then make based on the data, even if (as in the case of truncating by population) there is no apparent theoretical justification for not collecting complete data. For small-n or regional studies, statements of generality perform a similar function. The foundational work on "Varieties of Capitalism" (Hall and Soskice 2001), for example, leaves opaque the universe of possible cases to which the theory could apply. Empirical studies largely focus on post-industrial former British colonies, northern Europe and Japan, although there is some mention of France, Italy, Spain, Greece, and Turkey as countries sharing some of these characteristics. It is unclear what coding rule would place Turkey in a population of advanced industrial democracies, but not Mexico, Israel, South Korea, or South Africa. Without general statements of the population of cases or the sampling strategy—to say nothing of the classification of countries like France and Greece as a theoretically empty residual category of "Mediterranean economies"—it is likely that analysts have ignored critical variation in advanced capitalist economies, or included cases who do not fit properly into the population.

3.4. Replicability

The standard of replicability can be summarized through the following thought experiment. Given an existing variable or dataset, would a new coder tasked to create the same variable or dataset be able to recreate it identically? Coding schemes based on rigorous theory, with clear coding rules and precise statements of empirical scope and generality, should be easy

to recreate. The standard of replicability, though, suggests that researchers move further in order to ensure that variables and datasets are replicable. This links the CS to King's (1995) "Replication Standard" as well as work by Lindley (2006) on the replicability of case study research. King puts replicability at the forefront of empirical social science research, but suggests providing final datasets and codes for analysis with only scant attention paid to how researchers create such data sets.[2] There are several steps that qualitative and quantitative researcher alike can take to ensure that their variables, datasets, and coding decisions are just as replicable as their analyses. These steps follow the requisites of theory, clarity, and generality defined above.

The discussion of clarity above noted the many difficult choices necessary to code variables such as political violence. Yet in the vast majority of cases, the final dataset presents only cases that the coding rules dictate inclusion in the dataset. This leaves analysts with no way to assess whether or not coding principles have been applied correctly: data is censored at only affirmative decisions. Difficult coding decisions as naming all instances of political violence suggest an important refinement to the vast majority of coding practices. In addition to making clear statements about social facts that are requisites for variable coding decisions, coders should make data available for non-included cases as well as included cases. In other words, *when in doubt, include the observation, even if it falls out of the sample*. In principle, this should be relatively costless. Continuing the discussion of political violence, the primary methodology in current research is newspaper counts (Beissinger 2002; Varshney 2002; Varshney et al. 2004). If one is painstakingly reading all newspaper articles for which violence exists, it should be straightforward to note all cases of violence, and then to select from these all those cases for which the coding rules for "political" violence have been fulfilled.

[2] In their responses to King, Peterson (1995) and Maisel (1995) address this issue.

Of particular concern for quantitative researchers is that they *include all the steps for new variables constructed from other variables*. If a researcher has employed real GDP in constant prices as a variable in his analysis, but derived this variable using the GDP deflator and GDP in current prices, he should include the precise steps through which the new variable was constructed—including the sources for each existing variable and, if performed in a statistical package, the code through which the new variable was generated. This may seem excessive for simple transformations, but in many quantitative applications the steps between original variables and final variables employed in an analysis are far more complex. Determining the fraction of a country's GDP that comes from fuel exports from the World Development Indicators online database, for example, requires obtaining fuel exports as a percentage of merchandise exports, a measure of merchandise exports in current US dollars, and then GDP in current US dollars.

Data sources are similarly a paramount concern. Many coding projects synthesize data from disparate secondary sources, and among these sources there may be substantial disagreement—was Pinochet a personalist dictator or a military dictator, or both (Geddes 1999)? Coding rules ideally solve these problems, but nevertheless depend on the accuracy of the social facts being analyzed. In addition to clear coding rules, wherever possible, *coders should include in their datasets all sources for their coding choices*.

This is equally true for quantitative and qualitative researchers, although qualitative researchers do earn a special dispensation in the case of sensitive field research. The idea that qualitative researchers should at least strive for replicability is not new. King (1995) notes briefly that replicability should indeed apply to qualitative research, and in her response, Golden (1995) expounds on this view. *The only instances where coders should refrain from including*

information about data sources are instances from field research when interviewee or co-participant confidentiality agreements exist. These may be explicit confidentiality agreements, but alternatively, confidentiality agreements exist implicitly through failure of researchers to secure informed consent.

4. Examples from Comparative Politics and International Relations

To assess the extent to which modern political science research deviates from the Coding Standard, I randomly selected thirty journal articles from the fifteen top political science journals published between 2000 and 2006. The randomization procedure ensures that I cannot have chosen these "case studies" based on any *ex ante* considerations of their coding practices or substantive conclusions—inferences drawn from these examples are unlikely to either overestimate or underestimate the extent of coding problems in modern political science. To select articles, I assigned each article in a journal a unique integer beginning with 1, and then used the `sample` function in the statistical package R to draw at random two numbers without replacement from the list. I repeated this for all fifteen journals, for a total of thirty articles. Stratification by journal ensures that journals that publish more frequently than others (*Comparative Political Studies*) are not overrepresented.[3]

The population of articles from which random articles were sampled was constructed in the following manner. The journals were selected as the top political science journals by total citations, according to Journal Citation Reports® through ISI Web of Knowledge as of March 2007. The database separates international relations from political science journals. I chose the top five international relations journals, and then the top five general interest political science

[3] All data for this analysis, the sample of papers analyzed, the coding decisions and explications, and the lists of papers from which I sampled from *American Political Science Review*, *American Journal of Political Science*, *British Journal of Political Science*, *Journal of Politics*, *Political Science Quarterly*, and *Political Research Quarterly*, are available online at http://pantheon.yale.edu/~tbp4/docs/coding.zip.

journals, followed by the top five exclusively comparative politics journals without a regional or substantive focus.

Among the top international relations journals, two (*American Journal of International Law* and *Foreign Affairs*) were excluded from the analysis due to their focus on non-disciplinary topics. The top five international relations journals were therefore *International Organization*, the *Journal of Conflict Resolution*, *World Politics*, *International Security*, and *International Studies Quarterly*.

The top political science journals included the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *Political Science Quarterly*, and *Political Research Quarterly*. This left five political science journals that focused largely on comparative politics: the *European Journal of Political Research*, *British Journal of Political Science*, *Comparative Political Studies*, *Comparative Politics*, and *Politics and Society*. A number of political science journals that ranked higher on total citations were excluded from the analysis because of limited empirical scope—either in their substantive research questions or their geographical focus. These include *Public Opinion Quarterly*, *Public Choice*, *Annals of the American Academy of Political Science*, *Journal of Democracy*, *Political Geography*, *Journal of Peace Research*, *The Nation*, *Political Psychology*, *New Left Review*, and the *Journal of Common Market Studies*.

I excluded from analysis all book reviews, editors' introductions (but not introductory research essays), correspondences, responses, rejoinders, rebuttals, methodological primers,[4] review essays,[5] and country summaries. Inclusion in the population followed publication date

[4] "Methodological primers" are articles that introduced or reviewed methodological techniques. All were found in the *American Journal of Political Science*, i.e. Box-Steffensmeier and Zorn (2001), which was not included.
[5] "Review essays" were those that included a phrase such as "A Review Essay" in the title," or which fell under the heading of "Review Essay" in a journal's table of contents. One article which did not fulfill these criteria but which

rather than journal number. For example, *Comparative Politics* begins new volumes in October, so articles from Volume 32, Issue 1 (October 1999) were not included. *International Security* and *Political Science Quarterly*, which publish a "winter" issue spanning calendar years, begin at Winter 1999/2000.

All articles in *Comparative Politics*, *Comparative Political Studies*, *European Journal of Political Research*, *International Organization*, *International Security*, *International Studies Quarterly*, *Journal of Conflict Resolution*, *Politics and Society*, and *World Politics* were included in the population. It is clear that not all articles in these journals are empirical social science research: journals such as *Journal of Conflict Resolution* often publish articles exclusively based on formal models, and journals such as *International Security* often contain normative or policy articles, and journals such as *Comparative Politics* occasionally contain purely conceptual pieces. Nonetheless, selecting randomly, only pieces containing empirical social science research were chosen. For the remaining five general political science journals and the *British Journal of Political Science*, I used the following nested coding rules to judge an article as being an example of empirical comparative politics or international relations research.

- any article that included a cross-national statistical analysis was included;
- of the remaining articles, any article with a sub-national statistical analysis outside of the United States was included;
- of the remaining articles, any article that contained a description of a data set constructed using political information from outside of the United States was included;
- of the remaining articles, any article that used a case study in a country other than the United States was included;
- of the remaining articles, any article that contained a case study of relations between the United States and another country was included. This includes studies of American trade and foreign policy preferences and institutions.

---

was selected at random for analysis, Kenworthy (2001), surveys measures of wage setting institutions. After further inspection, I excluded the article from the population, as the article's evaluation of measures was inconsistent with the other articles' focus on original empirical research. I therefore re-sampled from the *World Politics* articles, with the sample at 106 (107 articles in the population, minus one already selected).

These coding rules created the population of cases under consideration. Data on the population of articles appears in Table 1.

Table 1: *Journals*

Journal Name	Field	Articles in Population
American Journal of Political Science	General	148
American Political Science Review	General	85
British Journal of Political Science	Comparative	140
Comparative Political Studies	Comparative	299
Comparative Politics	Comparative	142
European Journal of Political Research	Comparative	244
International Organization	IR	200
International Security	IR	145
International Studies Quarterly	IR	213
Journal of Conflict Resolution	IR	269
Journal of Politics	General	109
Political Research Quarterly	General	68
Political Science Quarterly	General	93
Politics and Society	Comparative	134
World Politics	IR	107

While the empirical scope of these findings is restricted to widely-cited work on comparative politics and international relations, the generalizability of the coding rules is straightforward. The universe of possible coding examples includes all empirical social science research, including that published in books, in lesser-cited academic journals, in other years, or in other disciplines and subfields. Although this study does not document current coding practices in these other domains, one could clearly extend this analysis to other research in these areas. The inferences that we draw based on this sample must recognize this non-random sampling procedure. It is possible, but not provable with current data, that book authors follow the CS more strictly than journal authors, for they have more room to expand on their coding procedures. Similarly, it is possible that other disciplines and subfields are better (or worse) than

comparative politics and international relations, and it is possible that a selection of lesser-cited journals would find less (or more) attention to coding practices.

To assess the degree to which an article follows the CS, I consider each of the CS's four components in turn. First, I check whether or not an explicit or implicit theory of variables and their values exists (theory). I code a theory as existing if the authors define the variables explicitly, or if such definitions are implicit in the citations or the discussion. For quantitative studies that employ variables from cross-national statistical databases such as tax ratios, democracy, or war as variables, I code theory as existing without explicit definitions what taxes, democracy, or war are. Second, I check whether or not there are clear statements of coding rules for variables and their values (clarity). A statement is clear if it gives readers sufficient information to reproduce the analysis. Third, I check whether or not there exists a statement of generality that relates the sample to the population of cases (generality). This exists only in articles where the authors explicitly state the relationship between the sample and the population (i.e. the sample is the population, the sample is random, or the sample is not random and rather conforms to a logic of "diverse examples"). A study may also note that it has not collected data on the entire population, but not state the relationship between what is effectively its sample and the population. I do not check to see for justifications of temporal bounds on case selection, as not a single study under consideration discussed its lack of temporal generality. Fourth, I check whether or not authors include data sources (replication). Evidence of replicability includes a list of sources for the data and any transformations done to them.[6] For theory, clarity, and replicability, three values—all, some, or none—are possible, depending on how many of the variables in the empirical analysis follow the coding standard. For generality, three values with

[6] Because this is a study of *coding* rather than *replication*, merely including a replication dataset without sources does not fulfill this criterion.

different interpretations—present, present without discussion of the consequences between sample and population, and absent—are possible.

Given those criteria, each paper was given a score of 0, 1, or 2 on each of four variables. Summing them together, scores for each papers could therefore range between 0 and 8. I also divided the papers into categories of quantitative and qualitative, on the coding rule that any paper that presented a regression coefficient was a quantitative paper. Based on this rule, the sample included ten qualitative papers and twenty quantitative papers.

I present the results in Table 2. The first three rows of data display the empirical means of each category, and the final row contains the test statistics and p-values for a Student's t-test that mean scores for Qualitative papers are less than mean scores for Quantitative papers.

Table 2: *Average Scores, by Dimension and Type*

	Theory	**Clarity**	**Generality**	**Replicability**	**Total**
All *N* = 30	1.47	1.00	0.77	1.60	4.83
Qualitative *N* = 10	0.90	0.10	0.60	1.50	3.10
Quantitative *N* = 20	1.75	1.45	0.85	1.65	5.70
T-statistic	-3.5586	-5.3581	-0.9492	-0.6165	-4.3366
p-value (1-tailed)	*0.0007*	*0.0000*	*0.1753*	*0.2713*	*0.0001*

Several points are worth noting. The raw scores (not reported) range from 1 to 7, indicating that not a single paper fulfilled all of the criteria of the CS. Moreover, on average, papers scored less than 5 out of 8. Both qualitative and quantitative papers did a fairly good job of indicating their sources, and neither qualitative nor quantitative papers fared well in terms of specifying the relationship between the observations they coded and the theoretical population of observations. But differences between qualitative and quantitative papers were substantively large and significant in terms of theory and clarity. Qualitative papers, by and large, performed far poorly

in terms of specifying the theoretical bases of their variables, and in particular in specifying clear observable coding criteria for their coding choices.

Lambert (2000) shows the consequences of unspecified theoretical concepts in qualitative coding. His study documents the relationship between globalization and class compromises in contemporary Australia. Yet neither globalization nor class compromise is ever defined as a theoretical concept in his work. At various points, globalization appears to refer to a policy outlook of an incumbent government, while at other points, it seems to correspond to the actual flows of goods and services into and out of Australia. Similarly, class compromise might correspond to an unspoken agreement between workers and factory owners, to a low level of retrenchment, or even simply to a low level of income inequality—it is impossible to tell which theoretical concept the author means by "class compromise." The result is that despite a fascinating account of evolving labor conditions and outward orientations between the 1970s and 1990s in Australia, we cannot adjudicate what precisely has changed, nor can we know the evidence upon which we should focus to ascertain the size, pace, or causes of that change.

Another study, Christiansen and Pallesen (2001), is emblematic of the consequences of a lack of clarity in coding for qualitative research. Their study of public sector reform in Denmark clearly lays out theoretical perspectives on what public sector reform in Denmark could mean, classifying different types of public sector reform and defining each. But turning to their empirical work, there is simply no set of clear criteria by which to judge what values these variables take. One might infer the values from their discussion, but it is impossible to gauge neutrally the state of reforms at any one point in time, or the empirical facts that correspond to changes in public sector operating procedures that constitute evidence of reforms. While their

evidence is highly suggestive, unclear coding choices prevent us from being able to adjudicate their findings.

Despite their comparatively superior performance overall, quantitative papers fell short of the CS in many respects as well. Moving to generality, Hiskey (2003) spends a considerable effort carefully defining concepts and detailing his coding decisions. Thereafter, his paper proceeds to analyze 237 municipalities in the Mexican states of Jalisco and Michoacan. On this, the coding procedure suffers from a lack of generality. We cannot tell from the text, for instances, if these 237 municipalities are all municipalities in the two states, or just a portion of them. Moreover, we have some limited sense of purposive sampling of the two states of Jalisco and Michoacan as part of a most-similar research design, but we still have no indication of the relationship between these two states and the rest of Mexico, or to other post-authoritarian middle-income states. Our ability to draw inferences suffers accordingly.

Finally, on replicability, Ferrara and Herron (2005) present a detailed analysis of pre-election coordination in a number of elections around the world. Their theoretical justification for measures and coding procedures are exemplary, and they provide internet links to a replication archive. Yet their provide almost no information about the data sources for their key independent and dependent variables, meaning that we can recreate their analysis given their data, but we cannot ascertain if they correctly created their data set given their original sources.

Altogether, the findings from a random sample of thirty comparative politics and international relations articles indicate that coding practices in modern political science research fall well short of the CS. Yet these findings are suggestive rather than conclusive. Three points are worth stressing in this regard. First, my coding decisions for theory, clarity, and replicability in most quantitative studies are quite generous. As noted above, I do not require statements of

theory for variables measured from existing datasets such as Polity IV. Moreover, I do not check the original datasets themselves for their own coding clarity or replicability. This is the case despite well-founded criticisms of coding practices for Polity IV (Munck and Verkuilen 2002), Minorities at Risk (Minorities at Risk Project 2005: 8), cross-national databases of economic statistics (Herrera et al. 2007), and others. Given that every cross-national quantitative study I reviewed included information from databases such as these, it is likely that I have understated the inferential difficulties from quantitative results. It is still possible, though, to level the same criticism against qualitative studies that cite other qualitative sources.

Second, it seems likely that the types of questions that qualitative and quantitative scholars ask differ. Quantitative scholars may focus on questions for which well-established theory and straightforward coding practices already exist, making it easy for them to avoid the coding difficulties inherent in topics such as the transformation of labor relations or party ideology over time, which qualitative scholars tackle.

Finally, none of these results entail that the results of the empirical analyses are incorrect. Indeed, it is possible that all are correct, and that a more careful presentation of evidence and coding choices would only strengthen their findings. The results only point to disturbing lack of precision and verifiability in empirical political science research, one that prevents us from ever knowing, on the basis of the evidence presented, the extent to which we can trust the inferences that authors wish us to make.

5. Conclusion

Michael Ward (2002: 48), responding to Munck and Verkuilen's (2002) survey of measures of democracy and their critique of coding practices, writes

> I heartily agree with their conclusion, which I take to be, "We do seek to emphasize that the careful development of measures constitutes the foundation for efforts at drawing causal inference and is a critical task in itself" (p. 31). I just do not think that they have provided very much guidance on how specifically to proceed anew or to renovate our very substantial intellectual capital in these data.

This paper, I hope, has fulfilled Ward's desideratum of how to "proceed anew," and pushed the problem of good coding practices beyond democracy scores and regime typologies. The CS describes a systematic process which all qualitative and quantitative data should be collected and coded. It relies heavily on theory and transparency, in particular on forthright statements of theoretical positions, clear articulations of observable facts that lead to coding choices, and transparent disclosures of choices and measures. As shown in the analysis of modern journals research in comparative politics and international relations, no articles follow the prescriptions of the CS precisely, and the inferences drawn from much qualitative and quantitative political science research is hence impossible to adjudicate.

As a final exercise, it is worth considering how the discipline itself might foster better coding practices. It is possible, after all, that my findings reflect not authors' failure to adhere to the CS, but rather journal editors' preferences for analysis rather than procedure in the texts that they publish. ("We are pleased to accept your paper for publication, conditional upon a revision of the text that removes your discussion of your coding procedures in order to meet our space requirements.")

An easy innovation would be to suggest that authors submit complete explanations of their coding decisions and procedures. Doing so should present no difficulties for good scholarship, as scholars engaged in standard coding practices should be able to accomplish this fairly easily. Only scholars for whom coding decisions are ad hoc or non-transparent will have

difficulty articulating a straightforward coding procedure, and this is no different from the now-standard assertion that an author who cannot state the procedure through which he obtained his regression coefficients should not publish those coefficients. These "coding appendices" need not be published within articles or books themselves, but instead might be included as web appendices, much as is currently becoming standard for replication archives of quantitative scholarship. The investments required to do so would be quite low, as the existing infrastructure already exists in the form of the Inter-University Consortium for Political and Social Research for quantitative data, and the Economic and Social Data Service Qualidata website in the United Kingdom for qualitative data.[7] Additionally, many journals such as the *Journal of Conflict Resolution* already require replication datasets to be submitted upon publication. It would be simple to extend this replication standard to coding rules and coding appendices.

I should also note that this paper is not only an exercise in methodological criticism, it is hopefully also an example of best practices. Throughout the "empirical" section of this paper, I have strived to reproduce the CS by articulating a clear theory of what makes coding practices good, defining the clear observable rules that lead me to make coding decisions that I make when analyzing other empirical social science research, stating the generality of my coding choices in terms of the sampling strategy and the relation of my sample to the population of coding examples, and giving data sources and replication archives. My own experience with best practices in coding was instructive in reinforcing the challenges often inherent in making coding choices, and in particular with articulating clear coding rules. This is a lesson that social scientists eager to produce cumulative social research are well-advised to learn.

[7] It is worth noting that the vast majority of U.S.-based qualitative researchers appear to be unaware of this resource.

6. References

Adcock, Robert, and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95 (3):529-546.

Alvarez, Michael, José Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1996. Classifying Political Regimes. *Studies in Comparative International Development* 31 (2):3-36.

Beissinger, Mark R. 2002. *Nationalist Mobilization and the Collapse of the Soviet State*. New York: Cambridge University Press.

Boix, Carles. 1999. Setting the Rules of the Game: The Choice of Electoral Systems in Advanced Democracies. *American Political Science Review* 93 (3):609-624.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, Kenneth A., and Robert Jackman. 1989. Democracy, Stability, and Dichotomies. *American Sociological Review* 54:612-621.

Box-Steffensmeier, Janet M., and Christopher J. W. Zorn. 2001. Duration Models and Proportional Hazards in Political Science. *American Journal of Political Science* 45 (4):972-988.

Brady, Henry E., and David Collier. 2005. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Christensen, Jørgen Grønnegård, and Thomas Pallesen. 2001. Institutions, Distributional Concerns, and Public Sector Reform. *European Journal of Political Research* 39 (2):179-202.

Collier, David, and Robert Adcock. 1999. Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts. *Annual Review of Political Science* 2:537-565.

Collier, David, and Steven Levitsky. 1997. Democracy with Adjectives: Conceptual Innovation in Comparative Research. *World Politics* 49:430-451.

Collier, David, and James E. Mahon, Jr. 1993. Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis. *American Political Science Review* 87 (4):845-855.

Cusack, Thomas R., Torben Iversen, and David Soskice. 2007 forthcoming. Economic Interests and the Origins of Electoral Systems. *American Political Science Review* 101.

Dahl, Robert Alan. 1971. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Fearon, James, and David Laitin. 2004. Civil War Narratives. Paper prepared for the workshop The Economic Analysis of Conflict: Problems and Prospects, Social Science Research Council, April 19-20, 2004.

Ferrara, Federico, and Erik S. Herron. 2005. Going It Alone? Strategic Entry under Mixed Electoral Rules. *American Journal of Political Science* 49 (1):16-31.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York: Wiley.

Geddes, Barbara. 1999. Authoritarian Breakdown: Empirical Test of a Game-Theoretic Argument. Paper presented at the Annual Meeting of the American Political Science Association.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring, John. 2004. What is a Case Study and What is it Good For? *American Political Science Review* 98 (2):341-354.

---. 2007a. *Case Study Research: Principles and Practices*. New York: Cambridge University Press.
---. 2007b. Is There a (Viable) Crucial-Case Method? *Comparative Political Studies* 40 (3):231-253.
Goertz, Gary. 2005. *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.
Golden, Miriam. 1995. Replication and Non-Quantitative Research. *PS: Political Science and Politics* 28 (3):481-483.
Gurr, Ted Robert. 1990. POLITY II: Political Structures and Regime Change, 1800-1986. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.
Hall, Peter, and David Soskice. 2001. *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*. New York: Oxford University Press.
Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47 (1).
Herrera, Yoshiko M., Devesh Kapur, and Sogomon Tarontsi. 2007. Improving Data Quality? What Is To Be Done? *APSA-CP* 18 (1):25-28.
Hiskey, Jonathan T. 2003. Demand-Based Development and Local Electoral Environments in Mexico. *Comparative Politics* 36 (1):41-59.
Hopf, Theodore. 2007 forthcoming. The Limits of Interpreting Evidence. In *Political Knowledge And Social Inquiry*, edited by Richard Ned Lebow and Mark Irving Lichbach. New York: Palgrave.
Kenworthy, Lane. 2001. Wage-Setting Measures: A Survey and Assessment. *World Politics* 54 (1):57-98.
King, Gary. 1995. Replication, Replication. *PS: Political Science and Politics* 28 (3):444-452.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
Lambert, Rob. 2000. Globalization and the Erosion of Class Compromise in Contemporary Australia. *Politics and Society* 28 (1):93-118.
Lijphart, Arend. 1969. Consociational Democracy. *World Politics* 21 (2):207-225.
Lindley, Dan. 2006. Presentation on Replicability and Case Studies. Tempe, AZ: 2006 Winter Institute for Qualitative Research Methods.
Little, J. Rodrick, and Donald Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
Mahoney, James, and Dietrich Rueschemeyer. 2003. *Comparative Historical Analysis in the Social Sciences*. New York: Cambridge University Press.
Maisel, L. Sandy. 1995. On the Inapplicability and Inappropriateness of the Replication Standard. *PS: Political Science and Politics* 28 (3):467-470.
Marshall, Monty G., and Keith Jaggers. 2000. Polity IV Project. University of Maryland, College Park.
Minorities at Risk Project. 2005. Minorities at Risk Project Dataset Users Manual 030703. College Park, MD: Center for International Development and Conflict Management. Available online at http://www.cidcm.umd.edu/mar/margene/mar-codebook_040903.pdf [Accessed May 1, 2007].
Munck, Geraldo, and Jay Verkuilen. 2002. Conceptualizing and Measuring Democracy: Evaluating Alternative Indices. *Comparative Political Studies* 35 (1):5-34.

O'Donnell, Guillermo. 1973. *Modernization and Bureaucratic-Authoritarianism: Studies in South American Politics*. Berkeley: Institute for International Studies, University of California.
Peterson, M.J. 1995. Community and Individual Stakes in the Collection, Analysis, and Availability of Data. *PS: Political Science and Politics* 28 (3):462-464.
Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley-Interscience.
Rokkan, Stein. 1970. *Citizens, Elections, Parties: Approaches to the Comparative Study of the Process of Development*. Oslo: Universitetsforlaget.
Sartori, Giovanni. 1970. Concept Misformation in Comparative Politics. *American Political Science Review* 64 (4):1033-1053.
Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
Sekhon, Jasjeet. 2004. Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals. *Perspectives on Politics* 2 (2):281-293.
Tobin, James. 1958. Estimation of Relationships for Limited Dependent Variables. *Econometrica* 26 (1):24-36.
Varshney, Ashutosh. 2002. *Ethnic Conflict and Civic Life: Hindus and Muslims in India*. New Haven, CT: Yale University Press.
Varshney, Ashutosh, Rizal Panggabean, and Mohammad Zulfan Tadjoeddin. 2004. Patterns of Collective Violence in Indonesia (1990-2003). Jakarta: United Nations Support Facility for Indonesian Recovery - UNSFIR Working Paper - 04/03.
Ward, Michael D. 2002. Green Binders in Cyberspace: A Modest Proposal. *Comparative Political Studies* 35 (1):46-51.