6

Theory Building and Hypothesis Testing: Large-N Versus Small-N Research on Democratization

Michael Coppedge

Two general approaches have been commonly used in the study of comparative politics: a 'thick' approach based on small-N comparisons, including case studies, and a 'thin' approach based on large-N comparisons. In Chapter 4, I introduced this distinction between thick and thin approaches and, focusing on issues of conceptualization and measurement, I identified some trade-offs associated with these approaches. Here I extend the discussion by addressing issues of theory and testing. As I will seek to show, both approaches make significant contributions. Moreover, even though their primary contributions focus on different aspects of the overall research process—small-N comparisons are invaluable with regard to theory generation and large-N comparisons are indispensable for hypothesis testing—I also argue that small-N researchers have a role to play, alongside large-N researchers, in the testing of theories.

The analysis draws on examples from the literature on regime change and democratization. Regime change is one of the oldest topics in political science. Even Aristotle, in the sixth century BC, analyzed transitions among democracy, aristocracy, and tyranny. And, because democratization has been studied for such a long time and by so many scholars, it has been subjected to every approach or methodology imaginable. For the small-N examples, I will rely heavily on Latin American research because it is most familiar to me. Fortunately, this area has launched several of the most engaging themes into the broader comparative democratization debate. Thus, this literature on regimes, democratization and Latin America provides a suitable point of reference for the analysis of thick and thin approaches to comparative politics.

6.1. THEORY GENERATION: COMPLEXITY AND SMALL-N COMPARISONS

Every theoretical model in the social sciences has five parameters. First, every model pertains to a certain level of analysis—individual, group, national, world-systemic, or some intermediate gradation between these. Second, it has one or more dependent variables. Third, it has one or more explanatory variables. Fourth, it applies to a certain relevant universe of cases. And fifth, it applies to events or processes that take place during a certain period of time. We can refer to the definitions of each of these five parameters as possessing zero-order complexity because no relationships among parameters are involved. In the study of democratization, however, even at the zero order there is great leeway for defining what democracy is, how to measure it and any explanatory factors, which sample of countries is relevant for testing any given set of explanations, and the period of time to which such explanations apply. And this is just at the national level of analysis; with smaller or larger units of analysis, one would use completely different variables, cases, and time frames.

First-order complexity involves any causal relationship between any of these parameters and itself. These relationships include:

1. Causation bridging levels of analysis or (dis)aggregation;
2. Causal relationships among dependent variables, or endogeneity;
3. Interactions among independent variables;
4. Impacts of one time period on another, called lagged effects or temporal autocorrelation; and
5. The impact of one case on another, called diffusion or spatial autocorrelation.

Second-order complexity involves causal relationships between two *different* parameters. All hypotheses about an independent variable causing democracy (or democracy causing something else) are of this order; but so are various complications that could be introduced into a model. If the meaning of democracy varies over time or the best way to operationalize an independent variable depends on the world region, then one is dealing with this degree of complexity. Third-order complexity comes into play when there are plausible hypotheses relating *three* parameters. Most common among these are hypotheses that the relationship between the dependent variables and an independent variable is partly a function of time or place. A good example is the hypothesis that the impact of economic development on democratization depends on a country's world-system position (O'Donnell 1973; Bollen 1983;

Hadenius 1992; Burkhart and Lewis-Beck 1994). With fourth-order complexity, a causal relationship could be a function of both time *and* place (or level of analysis). This may sound far-fetched, but in small-N comparison such relationships are fairly commonly asserted—for example, the notion that increasing wealth has not favored democracy in the Arab oil-producing states since World War II (Karl 1997); or the claim that the United States has become more sincerely interested in promoting democracy in the Caribbean Basin since the end of the Cold War (Huntington 1991).

Orders of complexity can increase only so far. Eventually, one arrives at the extremely inelegant 'saturated' model that explains each outcome perfectly by providing different and unique explanations for each case. Laypersons who have not been socialized into social science know that the saturated model is the truth: every country *is* unique, history never repeats itself *exactly*, and every event is the product of a long and densely tangled chain of causation stretching back to the beginning of time. We political scientists know on some level that a true and complete explanation for the things that fascinate us would be impossibly complex. But we willfully ignore this disturbing fact and persist in our research. We are a community of eccentrics who share the delusion that politics is simpler than it appears. This is why our relatives roll their eyes when we get excited about our theories. Although I would be as delighted as any other political scientist to discover simple, elegant, and powerful explanations, I think the common sense of the layperson is correct: we must presume that politics is extremely complex, and the burden of proof rests on those who claim that it is not. The ideal approach to theory generation would therefore reflect the complexity of the world.

When assessed according to this key criterion, the strength of theorizing based on small-N comparisons is readily apparent. Indeed, if a small-N approach to theorizing is compared to a suitable alternative such as rational choice theorizing—it bears clarifying that a large-N approach is not primarily a method for *generating* theory but instead a method for *testing* theory—small-N comparisons are clearly superior in generating hypotheses that faithfully reflect the complexity of the real world. In the case-based Latin American literature, the conventional wisdom presumes that each wave of democratization is different, that each country has derived different lessons from its distinct political and economic history, that corporate actors vary greatly in power and tactics from country to country, and that both individual politicians and international actors can have a decisive impact on the outcome. This is the stuff of thick theory, and comparative politics as a whole benefits when a regional specialization generates such rich possibilities.

The superiority is especially great in bridging levels of analysis, because rational-choice theory is anchored at the individual level. That is, rational choice theorizing aspires to make predictions about larger groups, but only within very restrictive assumptions about the rules of the game and the preferences of the players. And, as a result, it is difficult to extrapolate from these small settings to macrophenomena like regime change. Indeed, Barbara Geddes (1997) has called on scholars to stop trying to theorize about 'big structures, large processes, and huge comparisons', such as democratization, for the time being. In contrast, region-specific, small-N comparison has powerfully influenced the democratization research agenda for decades.

Examples abound. Juan Linz's theorizing (1978) about the breakdown of democratic regimes described a detailed sequence of events—crisis, growing belief in the ineffectiveness of the democratic regime, overpromising by semi-loyal leaders, polarization of public opinion, irresponsible behavior by democratic elites, culminating in either breakdown or reequilibration. He saw each step as necessary but not sufficient for the next, and described various options available to elites at each stage, as well as structural and historical conditions that made certain options more or less likely. This was a theory that assumed endogeneity, aggregation across levels of analysis, and conditional interactions among causal factors. In turn, Guillermo O'Donnell and Philippe Schmitter (1986) bridged levels of analysis when they theorized about democratization at the national level as the outcome of strategic maneuvering among elites at the group or individual level; they contemplated endogeneity or path dependence when they asserted that political liberalization was a prerequisite for regime transition.

Ruth Collier and David Collier's *Shaping the Political Arena* (1991) identified four similar processes or periods—reform, incorporation, aftermath, and heritage—in eight cases but allowed them to start and end at different times in each country. It was particularly exacting in describing the nature of oligarchic states, organized labor, and political parties and in specifying how they interacted with one another, and with many other aspects of their political contexts in the twentieth century, to affect the course of democratization. Finally, case studies of democratization, such as those collected in the Larry Diamond et al. (1999) project, and dozens of country monographs, weave together social, economic, cultural, institutional, and often transnational causes into coherent, case-specific narratives. In sum, the hypotheses generated by this small-N, case-based literature constitute significant contributions by reflecting high-order, complex theorizing.

Other desiderata of theory include (*a*) universal scope; (*b*) clear, simple, and explicit assumptions; and (*c*) the potential to generate testable hypotheses derived from theory. And when assessed by these criteria, rational choice has

a clear advantage over small-N comparisons with regard to scope. That is, rational-choice theory aspires to universal scope by refraining from limiting its applicability to certain times and places: what is true for one committee is assumed to be true for all committees as long as the assumptions of the model are met. In contrast, small-N comparisons typically generate theory about certain times and places rather than the universe. Indeed, small-N assumptions may be so specific that they are difficult to apply to other cases without wrestling with difficult issues of cross-national comparability.

But when the other criteria are considered, the small-N methods do quite well. Rational-choice theory makes its assumptions simple and explicit, which makes it easy for other scholars to follow the logic of the theory and derive the consequences of modifying some assumptions. And due to its deductive method, it lends itself to the generation of lots of hypotheses, especially about eventual, stable patterns of collective behavior. Yet, in a different way, small-N comparisons deliver similar benefits. Because the assumptions in small-N theorizing are well tailored to the cases at hand, they can be exceptionally clear and explicit. Moreover, the thick concepts used in such theorizing makes it possible to spin off many hypotheses about the causes and consequences of specific events.

In conclusion, the contributions of small-N comparisons, and to a lesser extent of rational choice, to the generation of theory should be recognized. In contrast, large-N comparisons are largely irrelevant to this task. Yet when it comes to theory, this is only one side of the equation. Indeed, it is one thing to develop a theory and quite another to develop a theory that is true. Whether the theory comes from deductive reasoning or extrapolating from inductive learning, it amounts to little if it does not conform to the evidence. This is what testing is about and this is where large-N comparisons become relevant again.

6.2. HYPOTHESIS TESTING: ASSESSING AND GENERALIZING ABOUT COMPLEX RELATIONSHIPS

If one accepts that the job of social scientists is to disconfirm all plausible alternative hypotheses, which are myriad, then one must also accept that all approaches yield only a partial and conditional glimpse of the truth. Nevertheless, all approaches have some value because, as it is often said, the truth lies at the confluence of independent streams of evidence. Any method that helps us identify some of the many possible plausible hypotheses is useful, as is any method that combines theory and evidence to help us judge how plausible these hypotheses are. But this perspective also suggests a practical and realistic

standard for evaluating the utility of competing methodologies. For methods that are primarily concerned with empirical assessments, it is not enough for a method to document isolated empirical associations or regularities; and it is asking too much to expect incontrovertible proof of anything. The question that should be asked is, rather, what are the strengths and weaknesses of each approach in helping us render certain kinds of alternative hypotheses more plausible or less?

6.2.1. Strengths and Limitations of Small-N Comparisons

On first thought, one might say that complex hypotheses cannot be tested using small-N methods because of the 'many variables, small-N' dilemma. The more complex the hypothesis, the more variables are involved; therefore a case study or paired comparison seems to provide too few degrees of freedom to mount a respectable test. This cynicism is not fair, however, because in a case study or small-N comparison the units of analysis are not necessarily whole countries. Hypotheses about democratization do not have to be tested by examining associations between structural causes and macro-outcomes. In Gary King, Robert Keohane, and Sidney Verba's terminology (1994: 24), we increase confidence in our tests by maximizing the number of observable implications of the hypothesis: we brainstorm about things that must be true if our hypothesis is true, and systematically confirm or disconfirm them.

The rich variety of information available to comparativists with an area specialization makes this strategy ideal for them. In fact, it is what these scholars do best. For example, a scholar who suspects that Salvador Allende was overthrown in large part because he was a socialist can gather evidence to show that Allende claimed to be a socialist, that he proposed socialist policies, that these policies became law, that these laws adversely affected the economic interests of certain powerful actors, that some of these actors moved into opposition immediately after certain quintessentially socialist policies were announced or enacted, that Allende's rhetoric disturbed other actors, that these actors issued explicit public and private complaints about the socialist government and its policies, that representatives of some of these actors conspired together to overthrow the government, that actors who shared the president's socialist orientation did not participate in the conspiracy, that the opponents publicly and privately cheered the defeat of socialism after the overthrow, and so on. Much of this evidence could also disconfirm alternative hypotheses, such as the idea that Allende was overthrown because of US pressure despite strong domestic support. If it turns out that all of these observable implications are true, then the scholar could be quite confident of the hypothesis. In fact,

she would be justified in remaining confident of the hypothesis even if a macrocomparison showed that most elected socialist governments have not been overthrown, because she has already gathered superior evidence that failed to disconfirm the hypothesis in this case.

The longitudinal case study is simply the best research design available for testing hypotheses about the causes of specific events. In addition to maximizing opportunities to disconfirm observable implications, it does the best job of documenting the sequence of events, which is crucial for establishing the direction of causal influence. Moreover, it is unsurpassed in providing quasi-experimental control, because conditions that do not change from time 1 to time 2 are held constant, and every case is always far more similar to itself at a different time than it is to any other case. A longitudinal case study is the ultimate 'most similar systems' design. The closer together the time periods are, the tighter the control. In a study of a single case that examines change from month to month, week to week, or day to day, almost everything is held constant and scholars can often have great confidence in inferring causation between the small number of conditions that do change around the same time. Of course, any method can be applied poorly or well, so this method is no guarantee of a solid result. But *competent* small-N comparativists have every reason to be skeptical of conclusions from macrocomparisons that are inconsistent with their more solid understanding of a case.

These comparisons within cases are the true strength of small-N methods. The benefit of doing comparisons across a small number of cases has been greatly exaggerated, because it is in such comparisons that the 'many variables, small-N' trap snaps shut with a vengeance. Small-N comparisons that are purely cross-national simply afford too little control to rule out the very large number of plausible alternative hypotheses, with the result that such studies end up being suggestive at best, or inconclusive at worst. Fortunately, scholars carrying out small-N comparisons, consciously or not, usually rely on within-case comparisons for their important evidence, and this is why they remain convincing.

This approach has two severe limitations, however. First, it is extremely difficult to use it to generalize to other cases. Every additional case requires a repetition of the same meticulous process-tracing and data collection. To complicate matters further, the researcher usually becomes aware of other conditions that were taken for granted in the first case and now must be examined systematically in it and all additional cases. Generalization therefore introduces new complexity and increases the data demands almost exponentially, making comparative case studies unwieldy.

The second limitation of the case study is that it does not provide the leverage necessary to test hypotheses of the third order of complexity and beyond.

Such hypotheses usually involve hypotheticals, for which a single case can supply little data (beyond interviews in which actors speculate about what they would have done under other conditions). For example, would the Chilean military have intervened if Allende had been elected in 1993 rather than 1970? If a different Socialist leader had been president? If he was in Thailand rather than Chile? If Chile had a parliamentary system? Such hypotheses cannot be tested without some variation in these added explanatory factors, variation that one case often cannot provide.

Harry Eckstein's advocacy (1975) of 'crucial case studies' sustained hope that some generalizations could be based on a single case. He argued that there are sometimes cases in which a hypothesis *must* be true if the theory is true; if the hypothesis is false in such a case, then it is generally false. But this claim would hold only in a simple monocausal world in which the impact of one factor did not depend on any other factor. Such a situation must be demonstrated, not assumed. In a world of complex contingent causality, we must presume that there are no absolutely crucial cases, only suggestive ones: cases that would be crucial if there were no unspecified preconditions or intervening variables. 'Crucial' cases may therefore be quite useful for wounding the general plausibility of a hypothesis, but they cannot deliver a death blow.

In turn, Douglas Dion's argument (1998) that small-N studies can be quite useful for identifying or ruling out necessary conditions is mathematically sound but probably not very practical. First, it does not help at all with sufficient conditions (or combinations of conditions), which we cannot afford to neglect. Second, it applies only when one already knows that the condition of interest probably is necessary and that any alternative explanations are probably not true. Given the complexity and diversity of the world, few conditions can be close to necessary, and the chances that *some* alternative explanation is true are very high. Therefore, such an approach is not likely to tell us anything we do not know already, and it is most likely that it will tell us nothing at all.

To sum up, though small-N comparisons have some strengths not only with regard to theory generation, as discussed in Section 6.1, but also theory testing, it has serious limitations when it comes to hypothesis testing. Indeed, given that focusing on 'few variables' would run against the theoretical inclinations of small-N researchers and amount to burying our heads in the sand, the only real solution to the 'many variables, small-N' problem is 'many variables, large N'. Thus, large-N comparisons, which provide the degrees of freedom necessary to handle many variables and complex relationships, provide a more suitable means for assessing, and generalizing about, complex relationships.

6.2.2. Promises and Challenges of Large-N Comparisons

Large-N comparisons need not be quantitative, as the qualitative Boolean analysis recommended by Charles Ragin (1987) has many of the same strengths. However, Boolean analysis forces one to dichotomize all the variables, which sacrifices useful information and introduces arbitrary placement of classification cutpoints that can influence conclusions (Elkins 2000). It also dispenses with probability and tests of statistical significance, which are very useful for ruling out weak hypotheses and essential for excluding the possibility that some findings are due to chance. Another weakness of Boolean analysis is that it greatly increases the risk of chance associations, which exacerbate its tendency to turn up many equally well-fitting explanations for any outcome and no good way to choose among them (see e.g. Berg-Schlosser and De Meur 1994).

Moreover, quantitative methods are available that can easily handle categorical or ordinal data alongside continuous variables, and complex interactions as well, so there would be little reason to prefer qualitative methods if quantitative data were available and sound. This is a conclusion with which Ragin should agree, as his principal argument against statistical approximation of Boolean analysis is that 'most data-sets used by comparativists place serious constraints on statistical sophistication' (Ragin 1987: 67). He is correct to point out that regression estimates might not be possible or meaningful if one were to specify all the permutations of interaction terms, as Boolean analysis does (Ragin 1987: 64–7). However, it is not clear that being able to obtain many rough answers, an unknown number of which are produced by chance, is an improvement over admitting that no answer is obtainable. Besides, social scientists should not be testing every possible interaction in the first place; they should only test those that seem plausible in the light of theory. 'Testing' them all without theoretical guidance is the definition of capitalizing on chance. Many large-N studies today have enough observations to handle dozens of variables and interactions with ease. The only truly satisfactory solution is to improve the quality and quantity of data across the board.

Of course, not everyone seeks general knowledge. This is partly a matter of taste. Sir Isaiah Berlin (1953) once suggested that people are either foxes, who know many small things, or hedgehogs, who know one big thing. I think a better analogy for my purposes would contrast whales and octopuses. Both are renowned for their intelligence, but they use their intelligence in different ways. Whales come to know great swaths of the earth in their tours of the globe; they lack limbs that would allow them to experience objects first-hand; and their eyesight is too poor to perceive fine detail. They acquire a surface

knowledge of general things. Octopuses, in contrast, dwell in one place and use their fine eyesight and eight infinitely flexible arms to gain an intimate knowledge of local, specific things. (To buttress the analogy, there is the additional, although not apropos, parallel that octopuses are well equipped to blend into their surroundings, while whales are as conspicuous as creatures can be. However, I ask readers not to overinterpret the octopus' tendency to spread clouds of ink when threatened.) I do not wish to suggest that scholars who emulate the octopus should emulate the whale instead, or vice versa. Rather, my point is that each kind of knowledge is limited in its own way and that the most complete kind of knowledge would result from pooling both kinds.

Limiting a sample to Latin America, for example, is not purely a question of taste; it also limits and biases what one can learn. Within-region comparison is often defended as a way of 'controlling' for factors that the countries of the region have in common, but this practice deserves a closer look. Such 'controls' would be effective if there were zero variation on these factors. But in many cases there is in reality quite significant variation on these factors within the region. Latin American countries, for example, were penetrated by colonial powers to different degrees, they were settled in different ways, their standards of living vary by a factor of ten, their social structures are quite distinct, many aspects of their political culture are unique, their relations with the United States and their neighbors are very different, they have evolved a great diversity of party systems, and there is a wide range in the strength of secondary associations and the rule of law. Bolivia and Guatemala should not be assumed to be comparable in each of these respects to Chile and Uruguay; yet this is exactly the assumption that the defenders of within-region comparisons make if they do not control directly for all of these differences. Therefore, limiting a sample to Latin America does not really control for these allegedly common factors very well.

Another problem is that there may not be enough variation in any of these factors to make controlling for them feasible in a regional sample. Although there is variation, it is often variation within a smaller range than what could be found in a global sample, and this may make it impossible to detect relationships. That is, in a truncated range variance is higher, which makes significance levels lower. Some important relationships with democracy are probably only observable over a global range. Indeed, as I have shown elsewhere (Coppedge 1997*a*: 190), though a relationship between socioeconomic modernization and democracy can definitely be perceived on a global scale, such a relationship would not necessarily hold up within the narrower range of variation found in Latin America or, for that matter, in Western Europe or Sub-Saharan Africa.

The inability to control adequately for certain variables makes it difficult to draw correct inferences. Donna Lee Van Cott (2000) turned up a fine example when she observed that party-system institutionalization is strikingly lower in countries with large indigenous populations than it is in most other Latin American countries. Statistically, institutionalization is negatively correlated with the size of the indigenous population; but it is also associated with other variables that correlate with indigenous population, such as income inequality, and which suggest a very different causal process. This creates a dilemma: one can either omit one variable and attribute all the influence to the other or include both and report that, due to the small sample and minimal variation in indigenous population and inequality over time, it is impossible to determine which matters or how much.[1] Yet an obvious cost is unavoidable: limiting the sample to a region makes it impossible to draw inferences outside the region. Any conclusions drawn from a Latin American sample implicitly carry the small print, 'This applies to Latin America. Relationships corresponding to other regions of the world are unknown.'

For both reasons, a cross-regional sample would always be preferable if other things—conceptualization, measurement, model specification—were equal. In practice, they rarely are equal: concepts, operationalizations, and theories are usually thinner in large-N studies. However, this thinness is a practical problem, not one inherent in the approach. The chief obstacle to large-N comparison is the scarcity of appropriate data: indicators of a great variety of thick concepts corresponding to large numbers of countries, at several levels of analysis, over a long period of time, sampled at frequent intervals. If such data were easily available, there would be no reason to avoid large-N, cross-regional comparisons.

The fact that little high-quality quantitative data are available for large samples is the main reason that the potential for large-N comparisons to explain democratization has not been realized more fully. For decades, large-scale testing of hypotheses about democratization lagged behind the sophistication of theories of democratization. Even very early theories of democratization—Alexis de Tocqueville's, for example—contemplated a multifaceted process of change. But it was not until the 1980s that scholars possessed the data required for multivariate, time-series analyses of democratization.

In the meantime, they did the best they could with the data that were available. There was quite of bit of exploration of thin versions of a variety of hypotheses. The central hypothesis in the 1960s was that democracy is a product of 'modernization', which was measured by a long, familiar, and

[1] Van Cott overcame this dilemma through within-case comparisons over time, but it remains a good example of the dilemmas encountered in within-region, cross-national comparisons.

occasionally lampooned set of indicators—per capita energy consumption, literacy, school enrollments, urbanization, life expectancy, infant mortality, size of industrial workforce, newspaper circulation, and radio and television ownership. The principal conclusion of these analyses was that democracy is consistently associated with per capita energy consumption or (in later studies) per capita GNP or GDP, although the reasons for this association remain open for discussion (Jackman 1973; Rueschemeyer 1991; Diamond 1992). Large-N studies also explored associations between democracy and income inequality (Bollen and Jackman 1985*a*; Muller 1988; Przeworski et al. 1996), religion and language (Hannan and Carroll 1981; Lipset, Seong, and Torres 1993; Muller 1995), region or world-system position (Bollen 1983; Gonick and Rosh 1988; Muller 1995; Coppedge 1997*a*), state size (Brunk, Caldeira, and Lewis-Beck 1987), presidentialism, parliamentarism and party systems (Mainwaring 1993; Stepan and Skach 1993), and economic performance (Remmer 1996).

This research also steadily forged ahead into higher orders of complexity. The first studies consisted of cross-tabulations, correlations, and bivariate regressions, taking one independent variable at a time. The first multivariate analysis was Phillips Cutright's in 1963 (Cutright 1963), but nearly a decade passed before it became the norm to estimate the partial impact of several independent variables using multiple regression. In the early 1980s some researchers began exploring interactions between independent variables and fixed effects such as world-system, a third-order hypothesis (Bollen 1983). However, these models were simpler than those being entertained by Latin Americanists of the time. O'Donnell's model (1973) of bureaucratic authoritarianism, for example, was nonlinear, sensitive to cross-national variations and the historical-structural moment, and defined the nature of links between the national and international levels of analysis (see also Collier 1979*b*). One major advance in the quantitative literature came in 1988, when Edward Muller (1988: 59–61) made a distinction between factors that cause transitions to democracy and factors that help already-democratic regimes survive. But this distinction was anticipated in the meetings of the Wilson Center group, held between 1979 and 1981, that led to *Transitions from Authoritarian Rule* (O'Donnell, Schmitter, and Whitehead 1986).

However, all of these studies were cross-sectional due to the lack of a time-series indicator of democracy. It was only in the 1980s that Freedom House and Polity data became available for a sufficiently large number of years to permit annual time-series analysis. These indicators are increasingly used to model change within large numbers of countries, rather than assuming that cross-national differences were equivalent to change (Burkhart and Lewis-Beck 1994; Przeworski et al. 1996; Power and Gasiorowski 1997;

Brinks and Coppedge 2006). Time series represent a great step forward in control, because they make it possible to hold constant, even if crudely, all the unmeasured conditions in each country that do not change from one year to the next. They therefore give one more confidence in inferences about causation.

Today large-N analysis does not uniformly lag behind the sophistication of theories generated by small-N research. In some respects, the testing is, aside from its conceptual thinness, on par with the theory. The state of the art in quantitative research on democratization now involves statistical corrections for errors correlated across time and space—so-called panel-corrected standard errors using time-series data (Beck and Katz 1995).[2] In lay terms, this means that analysts adjust their standards for 'significant' effects for each country (or sometimes region) in the sample, and also take into account the high likelihood that every country's present level of democracy depends in part on its past levels of democracy. These are, in effect, statistical techniques for modeling functional equivalence and path dependence. These corrections are, in my opinion, inferior to explicit specification of whatever it is that causes country-specific deviations and inertia, but so are most theoretical musings on the topic.

In other respects, quantitative analysis has inspired scholars to take theory into unexplored territory. For example, Adam Przeworski and Fernando Limongi (1997) were the first to develop in a systematic way the argument that transitions to democracy and democratic breakdowns were fundamentally different processes. They also contributed the concept of 'regime life expectancy', which has fired the imagination of scholars on both sides of the qualitative–quantitative divide. Another group of scholars has begun to explore the notion of democratic diffusion. Although Dankwart Rustow (1970) and Samuel Huntington (1991) wrote about various possible types of transnational influences on democratization, quantitative scholars have found that 'democratic diffusion' can refer to a tremendous variety of causal paths (Starr 1991; O'Loughlin et al. 1998; Brinks and Coppedge 2006). In the course of testing for them, they have had to refine the theory in order to distinguish among neighbor effects, regional effects, and superpower effects; impacts on the probability of change, change versus stasis, the direction of change, and the magnitude of change; and change influenced by ideas, trade, investment, population movement, military pressure, and national reputations, many of which were not contemplated in smaller-N or qualitative research.

[2] For applications of this and similar methods, see Burkhart and Lewis-Beck (1994), Londregan and Poole (1996), Przeworski et al. (1996), and Power and Gasiorowski (1997).

However, the large-N literature still lags behind the theory in at least three important respects. First, the concepts employed and measured remain thin, and their thinness lessens the value of all of this literature. Second, none of the large-N literature really addresses theories that are cast at a subnational level of analysis, such as the very influential O'Donnell–Schmitter–Whitehead (1986) project. Large-N testing concerns the national, and occasionally international, levels of analysis, and it will continue to do so until subnational data are collected systematically—an enterprise that has barely begun. Finally, there are quite a few hypotheses about causes of democratization that have not yet been addressed in large-N research. Among them are US support for democracy or authoritarian governments (Blasier 1985; Lowenthal 1991), relations between the party in power and elite interests (Rueschemeyer, Stephens, and Stephens 1992), the mode of incorporation of the working class (Collier and Collier 1991), interactions with different historical periods, US military training (Stepan 1971; Loveman 1994), and elite strategies in response to crisis (Linz and Stepan 1978; O'Donnell and Schmitter 1986).

6.3. CONCLUSION: IMPROVING THE PROSPECTS OF EMPIRICAL ANALYSIS

We are far from creating all the rich data that would be needed to combine the best of the small-N and large-N approaches. But even if this synthesis is a far-off dream, it is useful to keep in mind even today that these two approaches are parts of a whole and that, as data collection improves, we can expect them to converge rather than diverge into entirely separate camps. In the meantime, it is essential to maintain the bridges between small-N and large-N research. If scholars in both camps communicated better, we could achieve a more efficient division of labor that would accelerate social scientific progress. Large-N researchers should specialize in explaining the large, most obvious variations found in big samples. These explanations would define the normal expected relationships, which would serve as a standard for identifying the smaller but more intriguing deviations from the norm—the outliers. These outliers are the most appropriate domain for case studies and small-N comparisons, as they require a specialized, labor-intensive, sifting of qualitative evidence that is feasible only in small samples.

This is merely a call for each approach to do what it does best—large-N to sketch the big picture, small-N to fill in the details; some to look through all the jigsaw puzzle pieces searching for corner and sidepieces for the frame, others to fit together the pieces with similar colors and patterns. Neither camp needs to demean the work of the other; both make useful contributions to the

big picture. Those who specialize in small-N studies should not take offense at a division of labor that assigns them the outliers. This is in part because the outliers are the most interesting and challenging pieces, the ones with the greatest potential to innovate and challenge old ways of thinking. But another reason for not taking offense is that we already choose outliers as case studies. The rule of thumb is to choose cases where the unexpected has happened—'the unexpected' being defined with reference to general, large-N knowledge. At present, such selections are often done without systematic prior research. It would be an improvement to select cases for close study guided by more rigorous and systematic research.

Perhaps a 'division of labor' is an unfortunate metaphor, because if large-N and small-N scholars are truly divided, we cannot learn from each other. Instead of a division of labor, what we need is 'overlapping labors,' which requires some scholars to do research of both types and hence act as bridges. More broadly, in addition, we must communicate across the divide by reading work and attending conference panels outside our areas, always keeping an open mind and treating each other with respect, and never giving up hope that we can actually straddle it, individually or collectively.