# Theory Building and Hypothesis Testing: Large- vs. Small-N Research on Democratization

Michael Coppedge
Kellogg Institute
Hesburgh Center
University of Notre Dame
Notre Dame, IN 46556
574-631-7036
coppedge.1@nd.edu

Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 25-27, 2002.

**Theory Building and Hypothesis Testing: Large- vs. Small-N Research on Democratization**

Some practitioners of different approaches to comparative politics have jealously disputed one another's claims to theoretical understanding. Area studies specialists have accused quantitative comparativists of either comparing the incomparable or quantifying the obvious. Statistical analysts have condescendingly thanked country experts for digging up the anecdotal evidence that only multi-country comparisons can can transmute into theory. Both complain about the lack of realism in rational-choice theory, yet formal theorists have tried to brand propositions in both large-sample and case-study approaches "atheoretical" because they are not integrated into a larger, systematic body of propositions. All these charges reflect a narrow view of theory. In reality, all three approaches make indispensable contributions to good theorizing.

In this paper I define three fundamental criteria for evaluating theorizing and testing, and I use these criteria to evaluate three approaches in comparative politics–formal theory, case studies and small-sample comparisons, and large-sample statistical analysis. My purpose in doing so is to remind the subfield of a broader view of theory, in which each approach has one unique strength and two weaknesses. From this broad perspective, our three main approaches can be seen as complementary. (There are related complementarities in conceptualization and measurement, which I discuss in another paper.) I conclude by suggesting ways of improving our theories by combining approaches in innovative ways. I illustrate the tradeoffs with examples from research on democratization, which has been studied so long and so variously that it affords examples of the strengths and weaknesses of every method.

## Criteria for Good Theory

An overview of criteria for good theorizing provides a good foundation for a comparison of the advantages and disadvantages of different approaches. In a literature too vast to summarize here, scholars have defined more than a dozen criteria for good theory.[1] However, I contend that three criteria are especially central because they are locked in a three-way tradeoff: generality, integration, and thickness.[2]

*Generality*

A general theory is one that is intended to apply to all relevant cases, both all those that have been observed and all that could be observed.[3] (A general theory must also be correct for all cases, but I leave this discussion to the section on testing, below.) Some scholars claim not to seek general knowledge and consider the goal of generalization a matter of taste. Sir Isaiah Berlin once suggested that people are either foxes, who know many small things, or hedgehogs, who know one big thing.[4] 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.

For decades, the proponents of "middle-range theory" have fostered the view that generalization is, at best, optional, and at worst, impossible and pernicious. This is false. Generality is an indispensable characteristic of theory. In the standard (nomological) view of what theories are, an explanation interprets an event or a tendency as a specific instance of universal laws.[5] If the laws are not universal, then there is no solid foundation for the explanation; the explanation itself then requires explanation, and *that* explanation requires explanation, and so on. The phenomenon is not explained until it is understood as a necessary consequence of laws recognized as universally true.

Because true universals are unattainable in social science today, practicality forces us to confine our theories to bounded times and places. We must take care not to bound our theories arbitrarily. But as long as we can presume that there is potentially a good theoretical reason for limiting a theory to, say, postwar Europe, even if the reason is vague or implicit, then we can treat the theory as provisionally valid, pending completion (and empirical confirmation). All actual theories in comparative politics are therefore incomplete and provisional. The admission that they are works in progress is not damning, because this is all that one can claim about any scientific theory. And however inadequate this solution may be, in principle we cannot dispense with the obligation to generalize. This does not mean that all of us must generalize, only that some of us must.

*Integration*

As noted, generalization to the entirety of observed reality is not enough. In order to explain, we must also generalize to what is unobserved and hypothetical.[6] This is necessary to reassure ourselves that future observations will not contradict the generalization, since we cannot support a counterfactual empirically. Theory follows a different logic: instead of saying that a law is generally true because we have *observed* it to be generally true, theory says that a law is generally true because it is necessarily entailed by *other* laws that are generally true. (Those other laws are in turn subject to the same standards of confirmation.) Thus a generalization must be *integrated* into a systematic set of other laws--a theory--in order to be truly general.

The basis for the systematic structure of a theory is often logic, but it can be other branches of mathematics as well, such as calculus, game theory, or probability theory. I believe that it can also be, and typically and unavoidably is, common sense: our own direct experience of the world. No elaborate proof is needed to show that money talks, that united organizations are stronger than divided ones, or that you can fool some of the people some of the time. These understandings of how the world works are less foolproof than mathematical or logical tools, but consciously or not, we rely on them all the time. For example, if a general calls for the overthrow of an elected president, we may not know exactly what will happen, but the range of possible consequences is actually quite small. The general may be forced to retire or sent overseas, other officers may rally around him, the U.S. ambassador will weigh in or one side or another, and so on; but we know the consequences will not include the discovery of a platinum mine, a major earthquake, or the outbreak of world peace and harmony. Our common sense guides the translation of theoretical symbols into meaningful referents (interpretive theory) and informs and constrains the range of possible causal connections (causal theory).

Whatever the nature of the system that links hypotheses together, this integration has two benefits. First, the greater the number of propositions that are linked together, the more hypotheses they can generate. One can derive more theorems from fifty axioms than from three. As a result, complex, well-integrated theories have many observable implications and are therefore potentially

more testable (if their empirical referents are clear). Second, theory makes the cumulation of knowledge possible. As John Gerring notes, "The proposition that sits by itself in a corner is likely to be dismissed as 'ad hoc,' or 'idiosyncratic.' It does not fit with our present understanding of the world. It refuses to cumulate."[7] The more ways in which a proposition meshes with other propositions, the richer our understanding becomes.

*Thickness*

Finally, theory must be thick. A "thick" theory is one that involves many intertwined variables with effects that are conditional on time and place. In real life, a host of variables could contribute to any given effect; these variables can interact with one another; they can be causally linked among themselves; the nature of the causes or the effects can be different in different places or at different times; the causal relationship itself can vary by time or place, or both; effects can also be causes; cases can affect each other; and causation can be transmitted after a lag. True, all models necessarily simplify reality, and these simplifications are valued for their elegance, insight, and usefulness; but all simplifications necessarily sacrifice some accuracy. Only a complex theory can begin to approximate the richness of reality.

Guided by our own experience in the world, we should presume that most of these complex possibilities could be true. The burden of proof lies with those who claim that politics is simpler than it appears to the lay public. Of course, theoretical thoroughness does not guarantee a grasp of the truth; any creative person could dream up ten complex theories that are wrong for every one that is right. But very few of the right theories are likely to be simple. We should not let a misguided commitment to parsimony blind us to the truth. We have to consider complex theories; the trick is to find the right ones.

All approaches in comparative politics are deficient in satisfying some requirements for theory. In fact, each of the three major approaches excels in a different respect and is deficient on the other two. This is why they are competing approaches. Formal theory integrates propositions into a larger web of theory but neglects generalization and is thin, not thick; large-sample statistical analysis establishes general empirical fit, but in thin and often rather ad-hoc ways; and case studies and small-sample comparisons generate thick knowedge that may not be generally true and tends to be only loosely integrated into a larger theory. The following sections elaborate on this evaluation of each approach according to all three criteria, with illustrations from the literature on democratization.

*Theory in Small-Sample Studies*

Case studies and small-sample comparisons sometimes have been dismissed as "merely" descriptive, anecdotal, historical, or journalistic, and therefore atheoretical. But the harshest critic of such studies would have to concede that they at least generate "facts." Facts may seem atheoretical, but they are not. In reality, we cannot even observe facts until we have a conceptual framework to help us make sense of what we perceive.[8] Such a framework is an interpretive theory that helps us identify what to observe, defines the relevant and meaningful characteristics of actors and institutions, and fills in the connections between action and reaction so that we can plausibly reconstruct events and processes. We are unconscious of much of this theory; we develop it and refine it by trial and error from birth onwards. If the test of a good theory is that it enables us to predict consequences, the common sense we develop is a great theory. With it, we successfully walk, talk, drive, work, parent, and invest, negotiating our way around a thousand daily challenges

throughout whole lifetimes. With the benefit of this commonsense understanding of how the world works, we feel that we can understand political events that we did not experience directly, if only someone will supply us with the crucial details.

The more descriptive case studies and small-sample comparisons consist of propositions that are integrated into this intuitive interpretive theory. The bulk of democratization research consists of case studies and small-sample (usually within-region) comparisons. Every transition to democracy in the past two decades has been thoroughly analyzed in several books and numerous articles. Some of the most influential books in the field have been compendia of case studies. [9] Scholars seeking a thorough knowledge of a particular transition, breakdown, or regime survival are practically cursed with a superabundance of information. Often such studies prefer specific concepts to general ones: Duma to parliament, Clinton to president, Chiapas to province; but such precision reflects not a simple interpretive theory, but a more elaborate one that captures some of the idiosyncracies of each case.

What is striking at this level is that we collectively know so much and disagree so little. Research of this type has created what is probably the most thorough understanding of specific democratic transitions, breakdown, and survival, and has done so for practically every country one could mention. These works, whether they are academic research in a British or anthropological tradition, current history, or journalistic analyses, do an excellent, often unsurpasssed, job of recounting events, identifying key actors and their motives, asssessing the strength of organizations, and tracing connections among causal forces. The authority of this work is such that we rarely argue about who the key players were, what the basic chronology was, who won and who lost, or similar questions. Ironically, the lack of controversy about these inferences diminishes the prestige of the scholars who make them. But the high degree of consensus around their work should make their accomplishment more impressive, not less so. All theories should be as convincing as these.

But these studies are just one pole of a continuum in small-sample research. At the other extreme, some small-sample studies interpret every specific actor, institution, trend, and situation as an instance of a general type. They take literally Przeworski and Teune's call to "replace proper names of social systems" with "the relevant variables."[10] The kind of theory generated by this type of research tends to have two characteristics. First, most of it is qualitative and categorical. The causal relationships it identifies link types to types and kinds to kinds rather than matching quantities or degrees. Relationships are hypothesized to be true or false, necessary or sufficient, rather than partially true, stronger or weaker, likely or iffy. This qualitative bent does not make this style of theorizing inferior; rather, it is merely different from a mathematical style.

Second, the theoretical propositions that emerge from these studies, if examined with care, turn out to possess a high order of complexity. The more faithfully a theory represents our complex world, the more complex it must be. (*How* faithfully, of course, is a question to be resolved by testing.) In the Latin American democratization 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.

For these two reasons, case and area studies have made many of the best-known and most original contributions to comparative political theory. Dependency theory germinated in a study of Argentina's terms of trade.[11] Consociational democracy was inspired by Lijphart's Dutch origins.[12] The debate about the impact of parliamentary and presidential constitutions began as an effort to

understand the fall of the Weimar Republic and its renewal was inspired by the success of the Spanish transition.[13]

The hypotheses generated by this literature have reflected high-order, complex theorizing.[14] Daniel Lerner's seminal work on modernization was a case study of Turkey that identified parallel trends in the complex processes of urbanization, secularization, and education.[15] Juan Linz's theorizing about the breakdown of democratic regimes described a detailed sequence of events–crisis, growing belief in the ineffectiveness of the democratic regime, overpromising by semiloyal 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.[16] O'Donnell and Schmitter 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.[17] Huntington's thesis that there are waves of democratization required a transnational causal process in addition to multiple domestic causes.[18] The Colliers' *Shaping the Political Arena* 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 20th century, to affect the course of democratization.[19] Case studies of democratization, such as those collected in the Diamond, Linz, and Lipset projects and dozens of country monographs, weave together social, economic, cultural, institutional, and often transnational causes into coherent, case-specific narratives.[20] This literature has been the source of most of what we think we understand about democratization.

Nevertheless, the small-sample approach has two weaknesses. First, although its propositions are integrated with theory, they are integrated more loosely. Loose integration has two consequences. One is that the facts can be used to support an embarrassing variety of theories. This happens because the question,"What is this a case of?" has many possible answers. The leap from specific to general can go in many different directions. What, for example, was Venezuela in 1989 a case of? Every theoretical framework suggests a different direction. To a progressive political economist, it was an oil-dependent economy;[21] to an institutionalist, it was a presidential partyarchy;[22] to a liberal political economist, a case of delayed structural adjustment;[23] to a student of labor, it was a corporatist system;[24] to a cultural theorist, a nation with unrealistic trust in a "magical state."[25] In reality, all of these labels may have been accurate. The point is that moving from the specific to the general forces us to describe our cases more selectively, and we make our selections so as to integrate the case into a larger body of theory.

The second consequence of loose theoretical integration is that it is less clear which tests would confirm or disconfirm the theory. Without rigorous logic or mathematical tools to generate hypotheses, there is no straightforward way to derive *necessary* implications: what *must* be true if the theory is true. In contrast to formal theory, the theories of small-sample analysis are less clear about their assumptions; they rely more on the tacit assumptions of common sense, which leads to conditional and vaguely probabilistic predictions, which are hard to falsify.

The second weakness of small-sample theories is that they are, by definition, not general.

These propositions (when they are explicitly integrated into a theory) merely assert generality; whether such assertions are empirically valid or not is a matter for large-sample testing to decide. Until the testing takes place, these are only general hypotheses, not generally confirmed theory. Replacing proper names with variables is indeed our goal, but generalizing is far harder than affixing general labels to particulars. It is one thing to call the United States a presidential democracy, but quite another to assert that what one observes in the United States is true of presidential democracies in general. The former is a description of one case; the latter is an inference about a population (all presidential democracies) from one case (the United States), which is not justified.

To summarize, case studies and small-sample comparisons yield a type of theory that is qualitatively thick and empirically well-grounded, and therefore plausible in bounded times and places; but also provisional, pending extension to more general samples; and often ambiguous in its theoretical implications, and therefore difficult to test decisively, especially beyond its original boundaries. It is, to caricature a bit, a soft kind of theory built on a hard foundation.

*Theory in Large-Sample Comparisons*

Many scholars tend to view large-sample, statistical research an exercise in testing only, rather than a source of theoretical innovation. But even though the original motivation for applying statistics to comparative politics may have been to test hypotheses generated by other methods, this kind of research actually does contribute to theory in distinct and novel ways. The mathematical tools used in hypothesis testing encourage, and sometimes require, conversion of theories from a qualitative logic to a quantitative logic. Theories become less about kinds and types and true/false or necessary/sufficient relations and more about the magnitudes of impacts, partial impacts, probabilities, nonlinear relationships, and extrapolations beyond initial boundaries. These relationships are difficult to handle in a qualitative idiom. The reverse is not true, fortunately. Statistical analysis can also handle the kinds of relationships found in qualitative theories, such as conditional relations, necessary or sufficient conditions, and the direction of causation.

Examples of distinctly quantitative theory abound in democratization research. The qualitative hypothesis that wealthy countries tend to be democracies has been converted into a rococo variety of quantitative hypotheses:

- the wealthier the country is, the more democratic it is;
- the wealthier the country is, the more democratic it is, but with logarithmically diminishing increases;
- the wealthier the country is, the more democratic it is, but with logarithmically diminishing increases and a dip at an intermediate level of wealth (the "N-curve" hypothesis);
- the wealthier the country is, the more democratic it is, except when economic growth temporarily worsens inequality, which undermines democracy;
- the wealthier the country is, the more democratic it is, although the impact of wealth is mediated by state size, which has an "inverted U" relationship with democracy;
- increasing wealth does not make countries become more democratic but improves the life expectancy of any regime;

and so on. Another line of research has begun to explore the notion of democratic diffusion. Although Rustow and Huntington 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.[26] 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 vs. 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-sample or qualitative research.

The principal advantage of the kind of theory that emerges from large-sample work is that it is relatively general, both in its aspirations and in its empirical grounding. The degree to which it is general varies depending on the coverage of the universe by the sample, of course, but it is by definition more general than small-sample research. Formal theory makes universalistic assumptions, which are even more general, but large-sample research has the advantage of making at least some assumptions that are empirically supported. (The assumptions of statistical analysis are rarely fully supported, such as the assumption of normally distributed, random errors. I will discuss the consequences of this problem in the section on testing.) For example, the most consistent finding in the large-sample statistical literature is that democracy is associated with high levels of economic development. The association is a rough one, not strong enough to predict small differences in demoracy or differences between any two cases with a high degree of certainty; but it remains a very general statement.

The two weaknesses of large-sample comparisons are thinness and loose theoretical integration. A "thin" proposition is a simple statement that assumes very little about the world and identifies an association between just two narrowly conceived phenomena, such as democracy and development. Both could be, and originally were, thick concepts that would require thick theory. But large-sample research typically has reduced the concept of democracy to a few easily-measured institutions--fair and competitive elections, some basic freedoms--that reflect just one dimension of democracy: Dahl's notion of "contestation." Similarly, "economic development" has been reduced in this research to per capita GNP, GDP, or energy consumption. Thin concepts make for thin theory. Although the bivariate relationship between thin development and thin democracy has undergone elaborate permutations in statistical testing, many other hypotheses about the causes of democracy have been neglected. None of the large-sample literature really addresses theories that are cast at a subnational level of analysis, such as the very influential O'Donnell-Schmitter-Whitehead project. Large-sample research 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. In addition, there are quite a few hypotheses about causes of democratization that have not yet been addressed in large-sample research. Among them are U.S. support for democracy or authoritarian governments,[27] relations between the party in power and elite interests,[28] the mode of incorporation of the working class,[29] interactions with different historical periods, U.S. military training,[30] and elite strategies in response to crisis.[31] In this sense, the large-N literature lags behind the theories developed in other approaches.

The second weakness is loose integration with a body of systematic theory. Mathematics is an extremely systematic tool, but by itself it has no political content, and the political content that has been inserted into the mathematical framework lacks system. Large-N theory consists of a handful of isolated, disembodied propositions. Each one by itself can generate many necessary implications, simply by plugging in different numbers. But there is no theory in the gaps between the propositions that would enable us to combine them to make predictions. For example, we know that rich countries tend to be democratic, and that countries tend to become more like their neighbors. But there is no overarching theory that would enable us to predict how democratic a country should be if it is poor and surrounded by democracies, or rich and surrounded by authoritarian regimes. Lacking such a theory, quantitative researchers tend simply to accept whatever estimates they obtain as reflections of the true weights of different causal factors; and these

estimates are then absorbed into the conventional wisdom about what matters. It is a fundamentally inductive process of discovery. The problem with induction, as Hume observed centuries ago, is that even if a hypothesis is confirmed in 10,000 cases, there is no guarantee that it will be true in the 10,001st case unless one has a very good theory that predicts that it must.

Most of the large-sample research on democratization has concerned two debates: one about whether the relationship between development and democracy is linear, logarithmic, or some sort of N-curve;[32] and one examining interactions between per capita GDP, economic inequality and democracy.[33] The lack of specificity in this area is so great that after nearly forty years of repeated confirmation, a group of scholars was able to make a compelling case that the association is spurious.[34] Similar criticisms could be leveled against the emerging evidence for democratic diffusion (the spread of democracy from country to country): we are pretty sure it is happening, but we have little theoretical understanding of the causal mechanisms.[35]

This quality of knowing things but not knowing why is what inspires the accusation that large-sample research leads to "empirical generalizations" but not to theory. Again, I consider such criticism to be based on excessively narrow criteria for theory, one that privileges integration over general empirical confirmation (and thickness). The propositions that spring from large-sample analysis may be thin and disjointed, but they are still contributions to theoretical understanding.

*Formal theory*

Formal theories achieve levels of integration that elude small- and large-sample research. Three standards for good theorizing often touted by rational-choice theorists are universal scope; clear, simple, and explicit assumptions; and the potential to generate testable hypotheses derived from theory. 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. Rational-choice theory also makes its assumptions simple and explicit, which makes it easier for other scholars to follow the logic of the theory and derive the consequences of modifying some assumptions. Due to its deductive method, it also lends itself to the generation of lots of hypotheses, especially about eventual, stable patterns of collective behavior. Because a whole community of scholars follows this logic and works within it, new research builds explicitly on what has gone before. Theory cumulates.

However, the theory that cumulates is far from thick. Rational-choice theories of democratic transitions, for example, narrow their focus to the final stage of the process in which a small set of elites is bargaining about whether to found a democratic regime or not.[36] This is an extremely thin view of democratization. A thick theory would assume less and attempt to explain more. A thick theory would not limit the identities of the actors or the options before them to a fixed menu of choices. Game-theoretic explanations do not tell us how to know which actors are the elites and how they got their seats at the bargaining table. They do not explain where the bargainers' preferences came from or how they might change. They do not account for the number of rounds of bargaining to begin with, or why democracy is one of the options on the table. A thicker theory would offer explanations (whether cultural, institutional, international, social, or economic) for at least some of these elements; formal theories simply assume them. Formal theory, as currently practiced, has difficulty developing thick explanations because it is anchored at the individual level. It 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. It is the mirror image of small-sample theory: a hard theory built on a soft base. It is difficult to extrapolate from these small settings to macro-phenomena like regime change. Indeed, Barbara Geddes has called on scholars to stop trying

to theorize about "big structures, large processes, and huge comparisons," such as democratization, for the time being.[37]

Formal theories have universalistic aspirations; in this sense, they are even more general than large-sample theories, which are confined to the cases actually observed. However, generality is not, or should not be, merely an assertion or assumption of generality. It should be accompanied by empirical confirmation. In this more useful sense, formal theories encounter two obstacles to true generality. First, if taken literally, they rarely apply to the real world. All hypotheses generated by formal theories are derived from a set of premises. It would not be fair to test the hypotheses in cases where the premises did not hold. Yet the premises are typically extreme oversimplifications of reality–that there is a small set of players who perform complex calculations of self-interest with the benefit of full information and no history of past interactions, etc. Such premises cannot be said to hold in real-world situations, and therefore any test of predictions derived from them could be ruled unfair.[38]

Second, unlike small- and large-sample methods, formal theory is not a method of testing; it is only a method of generating theory.[39] In a very strict sense, the predictions of formal theories by definition have no empirical confirmation. In a less strict sense, their predictions can be tested; but if they are, they are tested using some version of small- or large-sample research, so the latter provide the only assurances of empirical support. But even if formal theories are generously credited with any empirical support their predictions find, it must be confessed that testing is the weak suit of formal theorists. As Green and Shapiro argue, "a large proportion of the theoretical conjectures of rational choice theorists have not been tested empirically. Those tests that have been undertaken have either failed on their own terms or garnered support for other propositions that, on reflection, can only be characterized as banal: they do little more than state existing knowledge in rational choice terminology."[40] One could object that in practice, formal theorists constantly adjust their assumptions and predictions to what they observe in the world to make it as realistic and relevant as possible. But this back-door induction is so unsystematic that it is prone to all the errors and biases found in the least rigorous testing methods. For these reasons, the following section on testing includes no further discussion of formal theory.

Nevertheless, the systematic, interlocking, cumulative nature of formal theory is an essential quality of theory, just as thickness and generality are. Each approach has one strength and two weaknesses, which complement the strengths and weaknesses of the other two approaches. Formal theory is integrated, but thin and lacking in general empirical support; large-sample analysis is relatively general, but thin and theoretically disjointed; and case studies and small-sample research are admirably thick but drastically bounded and too loosely integrated with a systematic set of theoretical propositions. Improved efforts at theorization should explore ways to combine the virtues of different approaches, such as deriving hypotheses from realistic axioms, testing hypotheses inspired by cases in large samples, and working backwards from empirical generalizations to logically coherent theories. But that is the subject of a different essay. The remainder of this essay evaluates the contributions of large- and small-sample research on democratization to testing.

## Testing

In debates about the merits of one approach vs. another for testing, it is healthy to bear in mind that all contain gaping methodological holes. We social scientists never prove anything, not

even with our most sophisticated methods. Popper argued that the goal of science is not to prove a theory, but to disconfirm alternative hypotheses.[41] In a strict sense, our goal is to disconfirm *all* the alternative hypotheses. But no serious social scientist requires proof that, for example, space aliens have not been destabilizing democracies by poisoning their water supplies. In practice, therefore, we are content to disconfirm only the alternative hypotheses that are conventionally considered plausible by other social scientists. (Of course, if implausible hypotheses become plausible later, we are obliged to try to disconfirm them as well.) This convention lightens our burden tremendously because the vast majority of the hypotheses an imaginative person could dream up are implausible. But it leaves room for a still-overwhelming number of alternatives, for two reasons. First, different people find different things plausible. Some people are convinced by intimate personal knowledge of a case, others by sophisticated statistical tests. Second, as Lakatos argued, disconfirmation is no simple yes-or-no exercise. Every hypothesis is embedded in a web of theories, not the least of which is the interpretive theory used to gather evidence for the test.[42] The common--and legitimate--practice of revising the supporting theories to explain away an apparent disconfirmation further increases the number of plausible alternatives.

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 Karl Deutsch 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. 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 alternative hypotheses more plausible or less?

One requirement that transcends approaches is that the theory be falsifiable, that is, we must be able to imagine some hypothetical evidence that would reveal the theory to be false. Unfortunately, falsifiability cannot be taken for granted. In fact, some influential theories have not been falsifiable. The problem can arise if, as in the case of some formal theories, the assumptions are not concrete enough to make it possible for the researcher to identify cases that would constitute a fair test. A more common problem in small-sample theories is a lack of clarity about what the theory predicts. Many reviews of literature on democratization list a dozen or more factors that favor democracy. These checklists make a clear prediction when all the favorable factors are present or absent, but not when some are present but others are not, and most cases are usually in this mixed category. Finally, some nonfalsifiable theories unconsciously employ circular logic. For example, Higley and Burton argued that democratic regimes are stable when the elite is united, and break down when elites are divided.[43] But the authors also judged whether elites were united or not in part by whether the regime broke down, which reduced their theory to a near-tautology.

*Testing in Case Studies and Small-sample 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 King, Keohane, and Verba's terminology, 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.[44]

The rich variety of information available to comparativists with an area specialization makes this strategy ideal for them. In fact, it is what they do best. For example, a scholar who suspects that 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 U.S. 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 macro-comparison 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. The thickness of the case study maximizes opportunities to disconfirm observable implications. In addition, 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 macro-comparisons that are inconsistent with their more solid understanding of a case.

These are the virtues of small-sample testing in principle. In practice, three problems are common. The first is indeterminacy, the classic "many variables, small N" problem. This tends to be more of a problem when the evidence is not as thick as it should be, so that even in a within-case comparison, there are more variables than observations. The result is that many different explanations fit the available evidence; there is no way to rule some of them out, so they all seem to matter. In practice, how do scholars deal with this problem? Sometimes they interpret the evidence selectively, presenting the confirming evidence that they prefer for ideological or extraneous reasons. In the worst cases, they may even suppress disconfirming evidence, consciously or not. These practices amount to interpretations, not tests. There is always a danger of avoiding a fair test when theory development and theory testing take place simultaneously, because the same evidence that was used to develop the theory was also used to "test" it.

But let us suppose that the scholar is honest and doing her best to be objective. When the evidence supports her theory, she keeps it; when it contradicts the theory, she modifies the theory to bring theory and evidence into agreement. She does this many times, iteratively, until she has a rich, plausible theory that fits the available evidence perfectly. And let us suppose that she has gathered such thick evidence that there is only one plausible theory that fits all the available evidence. Even in this ideal small-N situation, there is still a second and a third methodological problem.

The second problem is that the focus on one country exaggerates the importance of factors that happened to change during that period in that country. These tend to be explanatory factors that vary within countries over time in the short run, such as leadership, personality, weather and natural disasters, short-term economic crises, and so on. The third problem is that the focus on one country blinds the researcher to other factors that did not vary, or changed very slowly, even if they might be important explanations of cross-national differences or historical change. Together, these biased estimates of importance could be called myopia: focusing attention on what is small, close, and quick at the expense of what is large, distant, and slow-moving.

Comparisons of a few cases offer an advantage over single-case studies in principle, but in practice they are worse. One alleged advantage of comparisons is that they call attention to causal factors that are taken for granted when examining a single case. However, every additional case requires a repetition of the same meticulous process-tracing and data collection that was performed in the original case. 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 all additional cases. Comparison therefore introduces new complexity and increases the data demands factorially, making comparative case studies unwieldy.

Another alleged advantage is that tests can be run with new data that did not originally suggest the hypotheses. However, this advantage is lost every time the researcher modifies the theory in the light of new evidence. It is difficult to resist the temptation to salvage a theory by noting that it works differently under x circumstances. Every modification requires new evidence for a fresh test, so the many variables soon outstrip the small N once again.

But let us suppose the researcher manages to collect as much evidence in each of a dozen cases that was originally collected in a good case study. Is the problem of myopia avoided? Not completely. 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, (arguably the most homogeneous world region) 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. For an illustration of this, see the scatterplot of actual vs. predicted polyarchy levels by world region in my article on "Modernization and Thresholds of Democracy": there is definitely a relationship that can be perceived on a global scale, but which would not necessarily hold up within the narrower range of variation found in Latin America (or Western Europe or Sub-Saharan Africa; but it is evident in the most diverse region, Asia and the Pacific).[45] What is large, distant, and slow-moving is still only dimly perceived.

The inability to control adequately for certain variables makes it difficult to draw correct inferences. Donna Lee Van Cott 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.[46] 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. (Van Cott overcame this dilemma through thick within-case comparisons over time, but it remains a good example of the dilemmas encountered in within-region, cross-national comparisons.)

Of course, the obvious cost remains: 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."

Harry Eckstein's advocacy 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.[47] 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. Douglas Dion's argument 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.[48] 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.

Ultimately, therefore, small-sample testing contains no solution for the many variables, small-N problem. Even when such testing is done honestly, meticulously, and thoroughly, inferences to any larger sample are inevitably biased toward causes that vary within the sample and biased against causes that vary only outside the sample. Such studies are still very much worth doing, especially because they are firmly grounded in the real world and their concepts and propositions are satisfyingly thick. Middle-range theories can flourish in isolation from other contexts, but scholars who develop these theories should never forget that their vision is probably myopic and that their

findings are not likely to be fully comparable to findings from other contexts.

*Testing in Large-Sample Statistical Comparisons*

Generalization and complex relationships are better tested by large-N comparisons, which provide the degrees of freedom necessary to handle many variables and complex relationships. The only real solution to the "many variables, small N" problem (given that focusing on "few variables" would amount to burying our heads in the sand) is "many variables, large N." These large-N comparisons need not be quantitative, as the qualitative Boolean analysis recommended by Charles Ragin has many of the same strengths.[49] However, Boolean analysis forces one to dichotomize all the variables, which sacrifices useful information and introduces arbitrary placement of classification cut points that can influence the conclusions.[50] 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.[51]

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."[52] 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.[53] 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.

Nevertheless, long time series are available for only a few of the major variables needed to test theories of democratization. The fact that little high-quality quantitative data is available for large samples is the main reason why the potential for large-N comparisons to explain democratization has not be 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 – 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. Large-sample quantitative research has always been data driven, that is, its research agenda has been dictated by the availability of data more than by the priorities of theory.

Scholars have tried to make the most of 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 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.[54] Large-N studies also explored associations between democracy and income inequality;[55] religion and language;[56] region or world-system position;[57] state size;[58] presidentialism, parliamentarism, and party systems;[59] and economic performance.[60]

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 Cutright's in 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.[61] However, these models were simpler than those being entertained by Latin Americanists of the time. O'Donnell's model 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.[62] One major advance in the quantitative literature came in 1985, when Edward Muller made a distinction between factors that cause transitions to democracy and factors that help already-democratic regimes survive. But this distinction was anticipated by the discussions of the Wilson Center group that led to *Transitions from Authoritarian Rule* (published in 1986 but based on discussions held from 1979 to 1981).

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.[63] 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–panel-corrected standard errors using time-series data.[64] 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.

The scarcity of relevant data produces a second weakness of large-sample testing. Regression analysis is based on the assumption that all the variance that is not explained by the variables in the model is random noise. In effect, this means that the model omits no variables that are correlated with the explanatory variables that *are* in the model. But because data are scarce, many variables that are probably quite important are always omitted, and the chance that all of these are uncorrelated with the variables in the model is virtually zero. If this assumption is violated, then estimates of the impact of the explanatory variables are biased and inefficient, which means that judgments about whether their impacts are large or small, negative or positive, clear or cloudy, may be wrong.[65] Again,

there is no solution for this problem other than collecting more and better data. In the meantime, the theory that survives large-sample testing should be regarded as provisional, pending more rigorous tests using fuller specifications. The symmetry with small-sample testing becomes most obvious here. Small-sample results are provisional, awaiting testing in other cases and times; large-sample results are provisional, awaiting testing of thicker hypotheses.

**Prospects**

It is essential for us to build bridges between small- and large-sample research, and between these two methods and formal theory. If scholars in both camps communicated better, we could achieve a more efficient division of labor that would accelerate social-scientific progress. Large-sample 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 specialized, labor-intensive, qualitative evidence-sifting that is feasible only in small samples. Formal theorists could work to make assumptions increasingly realistic so that they can connect with the kinds of propositions used in more empirical work.

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 imagine how the final picture will look, others to look through all the jigsaw puzzle pieces searching for corner and side pieces for the frame, and still others to fit together the pieces with similar colors and patterns. No camp needs to demean the work of the others; all 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 theory and 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- and small-sample scholars are truly divided, we cannot learn from each other. Instead of a "division of labor," what we need is "overlapping labors," which require some scholars who do research of both types–perhaps not cutting-edge on either side, but valuable as a bridge. 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.

## Notes

1. John Gerring, in one of the most comprehensive discussions of "what is a good theory?" mentions generality ("breadth"), integration ("analytical utility" or "logical economy"), and thickness ("depth"), but also lists specification, accuracy, precision, parsimony, innovation, intelligibility, and relevance. John Gerring, *Social Science Methodology: A Criterial Framework* (Cambridge University Press, 2001), pp. 89-117.

2. Gerring's "specification" and "accuracy" are more pertinent to testing, discussed later in this chapter. I discuss precision as a characteristic of conceptualization and measurement in another paper. Innovation and intelligibility are lesser methodological virtues, and I do not consider parsimony a virtue at all.

3. Distinguishing between relevant and irrelevant cases is a crucial consideration. The rule of thumb is that a theory should not be arbitrarily bounded. In comparative politics, a theory is arbitrarily bounded when its application is restricted to certain objects for reasons that are not integral to the theory. A theory that applies only to Asian countries for no potentially theorizable reason is arbitrarily bounded, but one that explains how certain unique features of Asian countries condition the causal relationships is properly bounded.

4. Isaiah Berlin, *The Hedgehog and the Fox : An Essay on Tolstoy's View of History* (London: Weidenfeld & Nicolson, 1953).

5. Donald Moon writes, "The nomological pattern of explanation, as its name implies, requires the presence of general laws in any explanatory account. . . . But not just any kind of general statement can perform this explanatory function. Not only must laws be unrestricted universals (i.e., they must be in universal form and must apply to an unrestricted class of objects), but they must also support "counter-to-fact" and subjunctive conditional statements. . . . But to make such an assertion requires a great deal more information than that conveyed by a particular law, and so in order to understand the explanatory force of laws, we had to examine them in relation to scientific theories" (pp. 153-4). J. Donald Moon, "The Logic of Political Inquiry: A Synthesis of Opposed Perspectives," in Fred Greenstein and Nelson Polsby, eds., *Political Science: Scope and Theory*, vol. 1 of the *Handbook of Political Science*, pp. 153-54.

6."[The] 'orthodox' view of theories. . . has been called the 'dual language' conception of theories. In this view there are two crucial components of a theory: a set of theoretical principles and a set of correspondence rules which link the theoretical principles to observations, thereby providing an indirect or partial empirical interpretation to the theoretical principles. A theory of this type is what is often called a hypothetico-deductive system." Moon, "The Logic of Political Inquiry, p. 143.

7.Gerring, *Social Science Methodology*, p. 107.

8. Imre Lakatos, *The Methodology of Scientific Research Programmes,* ed. John Worrall and Greg Currie (Cambridge,Eng.: Cambridge University Press, 1978).

9. Juan J. Linz and Alfred Stepan, *The Breakdown of Democratic Regimes* (Baltimore: Johns Hopkins University Press, 1978); Guillermo O'Donnell, Philippe C. Schmitter, and Laurence Whitehead, eds., *Transitions from Authoritarian Rule: Comparative Perspectives*, (Baltimore: The Johns Hopkins University Press, 1986); Larry Diamond, Juan J. Linz, and Seymour Martin Lipset, eds., *Democracy in Developing Countries* 4 vols. (Boulder: Lynne Rienner, 1988-1989); Jorge Domínguez and Abraham Lowenthal, eds., *Constructing Democratic Governance: Latin America and the Caribbean in the 1990s* (Baltimore: Johns Hopkins University Press, 1996).

10. Adam Przeworski and Henry Teune, *The Logic of Comparative Social Inquiry* (New York: John Wiley and Sons, 1970), chapter 1.

11. Raul Prebisch, "El desarrollo económico de la América Latina y algunos de sus principales problemas" *El Trimestre Economico* 35:137 (1949); Kathryn Sikkink, "The Influence of Raúl Prebisch on Economic Policy Making in Argentina 1955-1962," *Latin American Research Review* 23:2 (1988): 91-114.

12. Arend Lijphart, *Democracy in Plural Societies: A Comparative Exploration* (New Haven: Yale University Press, 1977).

13. Ferdinand Hermens, *Democracy or Anarchy? A Study of Proportional Representation*, 2nd ed. (New York: Johnson Reprint Corporation, 1972); Juan Linz, "Presidential or Parliamentary Democracy: Does It Make a Difference?" in Juan Linz and Arturo Valenzuela, eds., *The Failure of Presidential Democracy* (Baltimore: The Johns Hopkins University Press, 1994), 3-87, but especially pp. 22-24.

14. The seminal article by Dankwart Rustow, "Transitions to Democracy" *Comparative Politics* 2 (1970): 337-63, anticipated many of the complex relationships discussed below.

15. Daniel Lerner, *The Passing of Traditional Society* (New York: Free Press, 1958).

16. Juan J. Linz, *The Breakdown of Democratic Regimes: Crisis, Breakdown, and Reequilibration* (Baltimore: The Johns Hopkins University Press, 1978).

17. Guillermo O'Donnell and Phillippe Schmitter, *Transitions from Authoritarian Rule: Tentative Conclusions about Uncertain Transitions* (Baltimore: The Johns Hopkins University Press, 1986).

18. Samuel Huntington, *The Third Wave: Democratization in the Late Twentieth Century* (Norman: University of Oklahoma Press, 1991).

19. Ruth Berins Collier and David Collier, *Shaping the Political Arena* (Princeton: Princeton University Press, 1991).

20. Diamond, Linz, and Lipset, eds., *Democracy in Developing Countries*.

21. Terry Lynn Karl, *The Paradox of Plenty: Oil Booms and Petro States* (Berkeley: University of California Press, 1997).

22. Michael Coppedge, *Strong Parties and Lame Ducks: Presidential Partyarchy and Factionalism in Venezuela* (Stanford: Stanford University Press, 1994).

23. Moisés Naím, *Paper Tigers and Minotaurs: The Politics of Venezuela's Economic Reforms* (Washington, DC: Carnegie Endowment for International Peace, 1993.)

24. Jennifer L. McCoy, "Labor and the State in a Party-Mediated Democracy: Institutional Change in Venezuela," *Latin American Research Review* 24 (1989): 35-67.

25.Fernando Coronil and Julie Skurski, "Dismembering and Remembering the Nation: The Semantics of Political Violence in Venezuela" *Comparative Studies in Society and History* 33:2 (April 1991): 288-337.

26.Harvey Starr, "Democratic Dominoes: Diffusion Approaches to the Spread of Democracy in the International System," *Journal of Conflict Resolution* 35:2 (June 1991): 356-381; John O'Loughlin, Michael D. Ward, Corey L. Lofdahl, Jordin S. Cohen, David S. Brown, David Reilly, Kristian S. Gleditsch, and Michael Shin, "The Diffusion of Democracy, 1946-1994," *The Annals of the Association of American Geographers* 88:4

(December 1998): 545-574; Daniel Brinks and Michael Coppedge, "Patterns of Diffusion in the Third Wave of Democratization," paper presented at the Annual Meeting of the American Political Science Association, Atlanta, September 2-5, 1999.

27. Cole Blasier, *The Hovering Giant: U.S. Responses to Revolutionary Change in Latin America, 1910-1985,* rev. ed. (Pittsburgh: University of Pittsburgh Press, 1985); Abraham Lowenthal, "The U.S. and Latin American Democracy: Learning from History," in Lowenthal, ed., *Exporting Democracy: The United States and Latin America,* pp. 261-83 (Baltimore: Johns Hopkins University Press, 1991).

28. Dietrich Rueschemeyer, John D. Stephens, and Evelyne Huber Stephens, *Capitalist Development and Democracy* (Chicago: University of Chicago Press, 1992).

29. Collier and Collier, *Shaping the Political Arena.*

30. Alfred Stepan, *The Military in Politics: Changing Patterns in Brazil* (Princeton: Princeton University Press, 1971); Brian Loveman, "'Protected Democracies' and Military Guardianship: Political Transitions in Latin America, 1978-1993," *Journal of Inter-American Studies and World Affairs* 36 (Summer 1994): 105-189.

31.Linz and Stepan, *The Breakdown of Democratic Regimes*; O'Donnell and Schmitter, *Transitions from Authoritarian Rule*.

32.Robert W. Jackman, "On the Relation of Economic Development and Democratic Performance," *American Journal of Political Science* 17 (1973): 611-21; Guillermo O'Donnell, *Modernization and Bureaucratic-Authoritarianism: Studies in South American Politics* (Berkeley: Institute for International Studies, University of California, 1973); Seymour Martin Lipset, Kyoung-Ryung Seong, and John Charles Torres, "A Comparative Analysis of the Social Requisites of Democracy," *International Social Science Journal* 136 (May 1993): 155-75; Adam Przeworski and Fernando Limongi, "Modernization: Theories and Facts," *World Politics* 49:2 (January 1997): 155-183.

33.Kenneth Bollen and Robert Jackman, "Political Democracy and the Size Distribution of Income," *American Sociological Review* 50 (1985): 438-57; Edward N. Muller, "Democracy, Economic Development, and Income Inequality," *American Sociological Review* 53:2 (February 1988): 50-68; Ross E. Burkhart, "Comparative Democracy and Income Distribution: Shape and Direction of the Causal Arrow" *Journal of Politics* 59:1 (February 1997): 148-164.

34. Przeworski and Limongi, "Modernization: Theories and Facts." To be fair, the economic development hypothesis was originally embedded in the rather elaborate theory of modernization, which held that democratization was a trend parallel with development, urbanization, education, secularization, and industrialization. These trends were held to be causally linked in complex ways, although the nature of the actors driving the processes was not well specified. Lipset's seminal article developed various additional arguments about the relationship. Seymour Martin Lipset, "Some Social Requisites of Democracy: Economic Development and Political Legitimacy" *American Political Science Review* 53 (March 1959): 69-105. However, over the years, as the demands of testing thinned concepts, non-economic aspects of modernization encountered inconsistent confirmation, and the inevitability of modernization came into question, the theory was reduced to a simple, underspecified hypothesis.

35.Laurence Whitehead, "International Aspects of Democratization," in Guillermo O'Donnell, Philippe C. Schmitter, and Laurence Whitehead, eds., *Transitions from Authoritarian Rule: Comparative Perspectives* (Baltimore: The Johns Hopkins University Press, 1986), pp. 3-46; Harvey Starr, "Democratic Dominoes: Diffusion Approaches to the Spread of Democracy in the International System" *Journal of Conflict Resolution* 35:2 (June

1991): 356-381; Brinks and Coppedge, "Patterns of Diffusion in the Third Wave of Democratization." But see Zachary Elkins, "Designed by Diffusion: International Networks of Influence in Constitutional Reform," paper presented at the 2001 Annual Meeting of the American Political Science Association, San Francisco, August 30-September 3, 2001.

36.Josep M. Colomer, "Transitions by Agreement: Modeling the Spanish Way," *American Political Science Review* 85:4 (December 1991): 1283-1302; Adam Przeworski, *Democracy and the Market* (Cambridge University Press, 1991), chapter 2, "Transitions to Democracy," pp. 51-99; Youssef Cohen, *Radicals, Reformers, and Reactionaries: The Prisoner's Dilemma and the Collapse of Democracy in Latin America* (University of Chicago Press, 1994), pp. 1-75; Gretchen Caspar and Michelle M. Taylor, *Negotiating Democracy: Transitions from Authoritarian Rule* (University of Pittsburgh Press, 1996); Barry R. Weingast, "The Political Foundations of Democracy and the Rule of Law," *American Political Science Review* 91: 2 (June 1997): 245-63; and the special issue of the *Journal of Conflict Resolution* 43:2 (April 1999), especially Mark J.C. Crescenzi, "Violence and Uncertainty in Transitions," *Journal of Conflict Resolution* 43:2 (April 1999): 192-212.

37. Barbara Geddes, "Paradigms and Sandcastles: Research Design in Comparative Politics," *APSA-CP Newsletter* (Winter 1997), pp. 19-20.

38. Donald P. Green and Ian Shapiro, *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science* (New Haven: Yale University Press, 1994), 38-42.

39. Hans-Peter Blossfeld, "Macro-Sociology, Rational Choice Theory, and Time: A Theoretical Perspective on the Empirical Analysis of Social Processes," *European Sociological Review* 12:2 (September 1996): 181-206.

40.Green and Shapiro, *Pathologies of Rational Choice Theory*, 6.

41. Karl R. Popper, *The Logic of Scientific Discovery* (New York: Harper and Row, 1968). Of course, we also need to demonstrate that our hypotheses do fit the evidence. This is not always easy, but it is easier than ruling out all the other possibilities, so I emphasize disconfirmation here. One of the alternative hypotheses is that any association we discover is due to chance. For this reason, some scholars encourage us to avoid procedures that increase the probability of false positives, such as testing a hypothesis with the same sample that suggested it or engaging in exploratory specification searches or "mere curve-fitting." Some even find methodological virtue in procedures that are more likely to generate hypotheses that are *wrong*, i.e., logical deduction of the implications of simplistic assumptions. I consider this stance an over-reaction to the danger of chance associations. The counter-intuitiveness of a hypothesis should increase our skepticism and our insistence on thorough testing, not our confidence in thinly documented associations. There are better ways to guard against false positives: enlarging the sample, replication using different indicators, and testing other observable implications of the hypothesis.

42.Lakatos, *The Methodology of Scientific Research Programmes*, pp. 8-101.

43.John Higley and Michael G. Burton, "The Elite Variable in Democratic Transitions and Breakdowns" *American Sociological Review* 54 (February 1989): 17-32.

44. Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994), p. 24.

45. Michael Coppedge, "Modernization and Thresholds of Democracy: Evidence for a Common Path and Process," in Manus Midlarsky, ed., *Inequality, Democracy, and Economic Development*, pp. 177-201 (Cambridge UP, 1997).

46. Donna Lee Van Cott, "Party System Development and Indigenous Populations in Latin America: The Bolivian Case," *Party Politics* 6:2 (2000): 159-162.

47. Harry Eckstein, "Case Study and Theory in Political Science," in Fred Greenstein and Nelson Polsby, eds., *Strategies of Inquiry* (Reading, Mass.: Addison-Wesley, 1975), pp. 79-138.

48. Douglas Dion, "Evidence and Inference in the Comparative Case Study," *Comparative Politics* 30:2 (January 1998), pp. 127-145.

49. Charles Ragin, *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* (Berkeley: University of California Press, 1987).

50. Zachary Elkins, "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations," *American Journal of Political Science* 44:2 (April 2000): 287-294.

51. Eg., Dietrich Berg-Schlosser and Gisèle De Meur, "Conditions of Democracy in Interwar Europe: A Boolean Test of Major Hypotheses," *Comparative Politics* 26:3 (April 1994): 253-80.

52. Ragin, *The Comparative Method*, p. 67.

53. Ragin, *The Comparative Method*, pp. 64-67.

54. Robert W. Jackman, "On the Relation of Economic Development and Democratic Performance," *American Journal of Political Science* 17 (1973): 611-21; Dietrich Rueschemeyer, "Different Methods, Contradictory Results? Research on Development and Democracy," *International Journal of Comparative Sociology* 32:1-2 (1991): 9-38; Larry Diamond, "Economic Development and Democracy Reconsidered," in Gary Marks and Larry Diamond, eds., *Reexamining Democracy*, pp. 93-139 (Newbury Park: SAGE. 1992).

55. Bollen and Jackman, "Political Democracy and the Size Distribution of Income"; Muller, "Democracy, Economic Development, and Income Inequality"; Adam Przeworski, Michael Alvarez, José Antonio Cheibub, and Fernando Limongi, "What Makes Democracies Endure?" *Journal of Democracy* 7:1 (January 1996): 39-55.

56. Michael T. Hannan, and Glenn R. Carroll, "Dynamics of Formal Political Structure: An Event-History Analysis," *American Sociological Review* 46 (1981): 19-35; Lipset, Seong, and Torres, "A Comparative Analysis of the Social Requisites of Democracy"; Edward N. Muller, "Economic Determinants of Democracy," *American Sociological Review* 60:4 (December 1995): 966-82, and debate with Bollen and Jackman following on pp. 983-96.

57. Kenneth Bollen,"World System Position, Dependency, and Democracy: The Cross-National Evidence" *American Sociological Review* 48 (1983): 468-79; Lev S. Gonick and Robert M. Rosh, "The Structural Constraints of the World-Economy on National Political Development," *Comparative Political Studies* 21 (1988): 171-99; Muller, "Economic Determinants of Democracy"; Coppedge, "Modernization and Thresholds of Democracy."

58. Gregory C. Brunk, Gregory A. Caldeira, and Michael S. Lewis-Beck, "Capitalism, Socialism, and Democracy: An Empirical Inquiry," *European Journal of Political Research* 15 (1987): 459-70.

59. Alfred Stepan and Cindy Skach, "Constitutional Frameworks and Democratic Consolidation: Parliamentarism and Presidentialism," *World Politics* 46 (October 1993): 1-22; Scott Mainwaring

"Presidentialism, Multipartism, and Democracy: The Difficult Combination," *Comparative Political Studies* 26 (July 1993): 198-228.

60. Karen L. Remmer, "The Sustainability of Political Democracy: Lessons from South America." *Comparative Political Studies* 29:6 (December 1996): 611-34.

61. Bollen, "World System Position, Dependency, and Democracy."

62. O'Donnell, *Modernization and Bureaucratic-Authoritarianism*; David Collier, "Overview of the Bureaucratic-Authoritarian Model," in David Collier, ed., *The New Authoritarianism in Latin America,* pp. 19-32 (Princeton: Princeton University Press, 1979).

63. Ross E. Burkhart and Michael Lewis-Beck, "Comparative Democracy: The Economic Development Thesis," *American Political Science Review* 88:4 (December 1994): 903-910; John B. Londregan and Keith T. Poole, "Does High Income Promote Democracy?" *World Politics* 49:1 (October 1996): 1-30; Przeworski et al., "What Makes Democracies Endure?"; Timothy J. Power and Mark J. Gasiorowski, "Institutional Design and Democratic Consolidation in the Third World," *Comparative Political Studies* 30:2 (April1997): 123-55; Brinks and Coppedge, "Patterns of Diffusion in the Third Wave of Democratization."

64. Nathaniel Beck and Jonathan Katz, "What To Do (and Not To Do) with Time-Series Cross-Section Data," *American Political Science Review* 89:3 (September 1995): 634-647.

65.Other assumptions of Ordinary Least Squares are often violated as well, but this is not a distinctive weakness of large-sample quantitative analysis vis-a-vis small-sample methods, which cannot even address such issues. The idea of a normal distribution and equal variances is difficult to handle without mathematical tools. A discussion of how best to detect and correct for violations of the statistical model is best handled in technical statistical references and does not belong in this comparison of large- and small-N methods of testing.