Thickening Thin Concepts and Theories: Combining Large N and Small in Comparative Politics
Author(s): Michael Coppedge
Source: *Comparative Politics*, Vol. 31, No. 4 (Jul., 1999), pp. 465-476
Published by: Comparative Politics, Ph.D. Programs in Political Science, City University of New York
Stable URL: http://www.jstor.org/stable/422240
Accessed: 06-12-2016 14:53 UTC

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.org/stable/422240?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://about.jstor.org/terms

Comparative Politics, Ph.D. Programs in Political Science, City University of New York is collaborating with JSTOR to digitize, preserve and extend access to *Comparative Politics*

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

Research Note

Thickening Thin Concepts and Theories

Combining Large N and Small in Comparative Politics

Michael Coppedge

In recent controversies about comparative politics, it often seems as though there are only two approaches: the systematic logical deduction of universalistic theory and the more traditional case studies and small-N comparisons leading more inductively to middle range theory.[1] The purpose of this article is to situate large-N quantitative analysis in this controversy. Quantitative analysis has its weaknesses, but they could be counterbalanced by some real strengths in small-N analysis. And quantitative analysis has certain methodological advantages that help compensate for some of the weaknesses of small-N analysis. On the one hand, small-N analysis tends to develop "thick" (complex or multidimensional) concepts and theories that are well-suited for description and for making inferences about simple causation on a small scale or in a few cases, but thick concepts and theories are unwieldy in generalizing or rigorously testing complex hypotheses. On the other hand, quantitative analysis is justifiably criticized for its "thin" (reductionist or simplistic) concepts and theories, but it is the best method available for testing generalizations, especially generalizations about complex causal relationships.

Quantitative analysis has hardly begun to exploit its full potential in assimilating complex concepts and testing complex theories, largely due to data limitations. In order to realize its potential, scholars need to answer two key questions that arise at the intersection of small-N and quantitative analysis. Can thick concepts be translated into the thin format of quantitative data? And can the nuanced, conditional, complex, and contextualized hypotheses of small-N analysis be translated into quantitative models? I argue that the answer to both questions is yes in principle, but that in order to make these approaches complementary in practice we must collect different data and more data and do it more systematically and rigorously.

A Perspective on Methods

In debates about the merits of different approaches it is healthy to bear in mind that all contain gaping methodological holes. Social scientists never prove anything, not even with our most sophisticated methods. Popper argued that the goal of science is

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

not to prove a theory, but to disconfirm alternative hypotheses.[2] In a strict sense, our goal is to disconfirm all the alternative hypotheses. However, in practice we are content to disconfirm only the alternative hypotheses that are conventionally considered plausible by other social scientists. (Of course, if implausible hypotheses later become plausible, 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, and still others by the force of logical deduction. 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.[3] The common—and legitimate—practice of revising the supporting theories to explain away an apparent disconfirmation further increases the number of plausible alternatives. Therefore, in this article I define disconfirmation rather loosely, as inconsistency with the web of theories conventionally treated as the facts.

Multiplication of plausible alternative hypotheses is especially problematic for those who would like to explain the complex macro phenomena of politics, because the number of plausible alternatives is almost hopelessly large. One can intuitively grasp the magnitude of the challenge by surveying all the orders of complexity involved.

Every theoretical model in social science has five parameters. First, every model pertains to a certain level of analysis—individual, group, national, world-systemic, or some intermediate gradation. 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 necessarily involved. However, even at the zero order there is great leeway in defining a concept, measuring it and any explanatory factors, selecting a relevant sample of countries for testing a set of explanations, and defining the period of time to which the explanations apply. And this example applies just to 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 within any of these parameters. These relationships include causation bridging levels of analysis, or aggregation and disaggregation, causal relationships among dependent variables, or endogeneity, interactions among independent variables, impacts of one time period on another, called lagged effects or temporal autocorrelation, and the impact of one case on another, called diffusion or spatial autocorrelation. Second order complexity involves interactions between two different parameters. Hypotheses in which an

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

independent variable X causes a dependent variable Y (or Y causes something else) are second order, but so are various complications that could be introduced into a model. If the meaning of Y varies over time or the best way to operationalize X depends on the world region, then one is dealing with second order complexity. Third order complexity comes into play when there are plausible hypotheses involving interactions among three parameters. Most common are hypotheses that the relationship between X and the dependent variable Y_i is partly a function of time or place. With fourth order complexity the causal relationship could involve interactions with both time and place (or level of analysis). This order of complexity may sound farfetched, but in small-N comparisons such relationships are fairly commonly asserted, for example, that increasing wealth has not favored democracy in the Arab oil-producing states since the second world war[4] or that the U.S. has become more sincerely interested in promoting democracy in the Caribbean basin since the end of the cold war.[5]

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. Each country is unique; history never repeats itself exactly; and each 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 of the things that fascinate us would be impossibly complex, but we wilfully 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. 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.

From this admittedly perfectionist perspective, 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 or judge how plausible these hypotheses are is useful. But this perspective also suggests a practical and realistic standard in evaluating the utility of competing methodologies. For methods that are primarily concerned with empirical assessments, documentation of isolated empirical associations or regularities is not enough, and incontrovertible proof can not reasonably be expected.[6] The question that should be asked is, rather, what strengths and weaknesses of each approach help render certain kinds of alternative hypotheses more or less plausible. These basic strengths and weaknesses can be discussed in terms of thick concepts, thick theory, and bridges between levels of analysis, three desiderata for a theory of macro processes in politics.

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

Thick Concepts

In the empiricist's ideal world, theoretical concepts would be "thin": simple, clear, and objective. We would theorize exclusively about relatively straightforward things like voter turnout, union membership, and legislative turnover. But in reality much of the political theory we find interesting concerns some of the messiest concepts—power, participation, legitimacy, identity, development, accountability, stability, and democracy. Such concepts can be called "thick" for two reasons. First, they can not be reduced to a single indicator without losing some important part of their meaning. Development is more than just average wealth; stability is more than just the absence of coup attempts; and democracy is more than just having elections. Second, thick concepts are often multidimensional, because no aspect of the concept is reducible to any of the others. For example, the breadth of the suffrage and the competitiveness of elections are both aspects of democracy, but knowledge of one does not allow us to predict the other reliably. Social scientists may disagree about how thick concepts should be, but at present, and as long as interesting theories are couched in terms of messy concepts, those who wish to test such theories have no choice but to translate those concepts into indicators of one sort or another, whether the results are categorical definitions or continuous numerical variables.

Small-N analysis excels at the kind of conceptual fussiness that is required to develop valid measures of thick concepts. Researchers using this approach usually define their concepts carefully. They take pains to explain how their definition of a concept differs from their colleagues'; they spend a great deal of time justifying the functional equivalent of the concept in the case they are analyzing; they are sensitive to changes in the meaning of the concept over a long period of time; and they debate publicly what is or should be meant by the word that represents the concept. One demonstration of these tendencies is Collier and Levitsky's recent survey of qualifiers for "democracy": they encountered hundreds in the published literature![7] This attention to nuance comes at a price, however, for it impedes generalization and cumulation. The more elaborately a concept is defined, the narrower it becomes. The more baggage it has to carry, the less widely it can travel. This difficulty in generalizing also means that general theory cumulates accompanied by cumulative uncertainty. If my explanation of Y_1 differs from his explanation of Y_2 or her explanation of Y_3, we may be explaining slightly different things. Every researcher who defines a dependent variable anew automatically lends plausibility to this alternative hypothesis, which remains plausible until it is ruled out by additional research.

Quantitative research has the opposite strengths and weaknesses. Its variables tend to be defined more narrowly, which makes it more feasible to gather data from a large number of cases and therefore to support generalizations. Also, the same few variables tend to be used repeatedly. This habit, which is reinforced by the cost of collecting new data for a large sample, reduces (but does not eliminate) the plausi-

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

bility of the hypothesis that different researchers are in fact explaining different things in different ways and therefore favors (but does not guarantee) cumulation. However, the thin concepts implied by the construction of some of the variables often introduces uncertainty about the validity of these measures. Quantitative researchers in effect use the bait-and-switch tactic of announcing that they are testing hypotheses about the impact of, for example, economic development and then by sleight of hand substitute an indicator of per capita energy consumption and assert that it measures development well enough. The problem with such substitutions is not necessarily that they do not measure the concept of interest at all, but that a single narrow indicator can not capture all the relevant aspects of a thick concept.

It is tempting to hide behind the excuse that we often test only the implications of the hypothesis rather than the hypothesis itself. But this fig leaf offers no real protection, because a valid test of the full hypothesis would still require the testing of all of its manifold implications. Often these tests would be the same as tests involving an equivalent complex multidimensional indicator. For example, if theory tells us that X causes Y, it makes little difference whether we treat Y's thin indicators Y1 and Y2 as observable implications of Y or as component dimensions of Y; either way, we end up testing for associations between X, on the one hand, and Y1 and Y2, on the other. Unpacking a thick concept, exploring its dimensionality, and translating it into quantitative indicators can be seen as a process of discovering more of the observable implications of a theory and therefore of rendering it more testable.

Is there any way to assemble large datasets with valid indicators? It would be easier if concepts were not thick. But some thick concepts are too old and too central to our thinking to be reduced in this way. In such cases the alternative is to explore empirically how to measure them validly and reliably, which requires recognition of their complexity. A basic procedure in measuring any complex concept has four steps. First, the analyst breaks the mother concept up into as many simple and relatively objective components as possible. Second, each of these components is measured separately. Third, the analyst examines the strength of association among the components to discover how many dimensions are represented among them and in the mother concept. Fourth, components that are very strongly associated with one another are treated as unidimensional, that is, as all measuring the same underlying dimension, and may be combined. Any other components or clusters of components are treated as indicators of different dimensions. If the mother concept turns out to be multidimensional, the analyst then has two or more unidimensional indicators that together can capture its complexity. If the mother concept turns out to be unidimensional, then the analyst has several closely associated component indicators that may be combined into a single indicator that captures all the aspects of that dimension better than any one component would.[8]

Controversy has always surrounded quantitative indicators. One basic objection holds that, when the theoretical concept is categorical, not continuous, then attempts

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

to measure it with a continuous instrument produce either measurement error or nonsense.[9] But if both continuous and categorical indicators measured exactly the same concept, then we would prefer the continuous one on the grounds that it is more informative, more flexible, and better suited for the sophisticated testing that can rule out more of the plausible alternative hypotheses. If one wanted a categorical measure, it could always be derived from the continuous one by identifying one or more thresholds that correspond to the desired categories. A dichotomized indicator would sort cases and interact with other variables the same way a dichotomy would, again assuming that they measured exactly the same concept. In other words, the continuous indicator contains more information, which we could choose to ignore, but the reverse is not true: one can not derive a continuous measure from a categorical one without adding a great deal of new information.

Some may still object that the additional information in the continuous measure is not meaningful or useful because translation of neat and satisfying categories into slippery matters of degree deprives us of analytic footholds. According to this argument, our minds seek out categories because we need definite, firm, satisfying, categorical ideas to guide us. This argument, however, is just an illusion created by attempts to translate precise mathematical language into imprecise verbal language. Let us suppose that a simple bivariate regression estimate of the relationship between democracy and per capita GNP is Democracy = 1.5 + 1.33*log(GNPPC). If one had to explain this finding without using numbers, one could say little more than there is a minimal level of democracy below which no country falls, but the wealthier the average person is, the more democratic the country is. However, the benefits of wealth diminish steadily as wealth increases. Such a statement does not allow one to make any useful statement about how democratic we should expect any country to be. But this statement is only a faint hint of what the estimate really says because the most useful information—the numbers—have been removed. Restoring the numbers recreates a compact, elegant formula that can generate quite definite predictions. For example, if per capita GNP were $3,000, this formula predicts a democracy score of 6.12. Properly understood, it is not a false precision, because the standard errors and other parameters of the estimate can be used to calculate a confidence interval for any of its predictions.

It is of course natural to feel uncertain about what a prediction of 6.12 means because the number itself has no inherent meaning. But the same could be said about other numbers in our everyday lives—temperature readings, ages, incomes, and Olympic diving scores. All of these numbers, and the equally arbitrary words that make up a language, acquire meaning only through the familiarity that comes from using them to describe, compare, and communicate. Theorists could certainly refuse to recognize the higher and lower ranges of a continuous indicator as valid, on the grounds that they were never contemplated in the original theory. But scholars who define those higher and lower ranges are breaking new conceptual ground. There is

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

no reason not to develop and use thick continuous measures as long as one threshold of their continuous concept is faithful to all the facets of the original categorical concept.

This defense of the superiority of continuous indicators rests entirely on the premise that the hypothetical continuous and categorical indicators measure exactly the same concept. In theory they could, but in practice they often do not. The problem with many quantitative indicators is not that they are quantitative, but that they are qualitatively different from the categorical concepts they purport to measure. The real problem with continuous indicators is that they measure only thin, reductionist versions of the thicker concepts that interest nonquantitative scholars.

Thick Theory

A second strength of small-N comparison is the development of thick theory: richly specified, complex models that are sensitive to variations by time and place. As argued above, such complex models are desirable because many of the complex alternative hypotheses are plausible and we must try to disconfirm them in order to make progress. In the study of complex macro phenomena the virtues of parsimony are overrated. Small-N comparative work does a good job of suggesting what these complex relationships might be. In the small-N literature the conventional middle range wisdom presumes that generalizations are possible only within carefully circumscribed historical periods, 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 outcomes. This is the stuff of thick theory, and comparative politics as a whole benefits when a regional specialization generates such rich possibilities.

Can such complex hypotheses be tested with small-N comparisons? On first thought, one might say no because of the "many variables, small N" dilemma. The more complex the hypothesis is, 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. We increase confidence in our tests by brainstorming about things that should be true if our hypothesis is true and systematically confirming or disconfirming them.[10] 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

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

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, and that the opponents publicly and privately cheered the defeat of socialism after the overthrow. Much of this evidence could also disconfirm alternative hypotheses, for example, 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. In addition to maximizing opportunities to disconfirm observable implications, it does the best job of documenting the sequence of events, which is crucial in establishing the direction of causal influence. Moreover, it is the ultimate "most similar systems" design, 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. The closer together the time periods are, the tighter the control is. In a study of a single case that examines change from month to month 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. Competent small-N comparativists have every reason to be skeptical of conclusions from macro comparisons that clash with their more solid understanding of a case.

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 exponentially, making comparative case studies unwieldy. Second, the case study does not provide the leverage necessary to test counterfactual hypotheses, 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

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

rather than 1970? If a different Socialist leader had been president? If he were in Thailand rather than Chile? If Chile had a parliamentary system? Such hypotheses can not be tested without some variation in these added explanatory factors, variation that one case rarely provides.[11]

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.[12] 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 in wounding the general plausibility of a hypothesis, but they can not deliver a death blow.

Generalization and complex relationships are better supported by large-N comparisons, which provide the degrees of freedom necessary to handle many variables and complex relationships. These comparisons need not be quantitative, as qualitative Boolean analysis has many of the same strengths.[13] However, Boolean analysis forces one to dichotomize all the variables, which sacrifices useful information and introduces often arbitrary placements of classification cut points that can influence the conclusions. It also dispenses with probability and tests of statistical significance, which are very useful in ruling out weak hypotheses. Moreover, quantitative methods that can easily handle categorical or ordinal data alongside continuous variables, and complex interactions as well, are available, so there would be little reason to prefer qualitative methods if quantitative data were available and sound.

The lack of high quality quantitative data for large samples is the main reason why the potential of large-N comparisons has not been realized more fully. Ideally, data collection would also be done more rigorously so that all the variables employed would be adequately valid indicators of the concepts of theoretical interest. This rigor is not always possible, but where it is not, as with data collected worldwide for other purposes, greater use should be made of statistical techniques for incorporating latent variables into models. These techniques can help compensate and correct for poor measurement if several related indicators are available and their theoretical relationship to the concept of interest is known.

Bridging Levels of Analysis

Most of the quantitative research on macro phenomena is cast at the national level of analysis. The widest gulf that divides large-N studies from small-N comparisons

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

results from the fact that most of the latter are either cast at the subnational (group or individual) level or move easily between all levels, from individual to international. Small-N comparison is more flexible in this respect. Case studies routinely mix together national structural factors such as industrialization and constitutional provisions, group factors such as party or union characteristics, individual factors such as particular leaders' decisions, and international factors such as IMF influence. Quantitative researchers are caught flat-footed when faced with shifting levels of analysis because they go to great pains to build a dataset at one level. Incorporation of explanatory factors from a lower level requires their building a completely different dataset from scratch. Their units of analysis are countries and years, at best. For example, to test the hypotheses on regime transitions from the O'Donnell-Schmitter-Whitehead project, one would have to collect data about strategic actors rather than countries and resample at intervals of weeks or months rather than years.

In view of the difficulty of bridging levels of analysis, it is tempting to conclude that the effort is not necessary, that the choice of a level of analysis is a matter of taste and those working at the individual and national levels may eat at Almond's separate tables and need never reconcile their theories. But from the perspective of methodological perfection outlined in this article, the level of analysis is not a matter of taste, because no level of analysis by itself can yield a complete picture of all the causal relationships that lead to a macro outcome. All levels of analysis are, by themselves, incomplete. Rational choice advocates are right to insist that a political theory tell us what is going on at the individual level. However, it is wrong to argue that associations at a macro level do not qualify as theory until one can explain the causal mechanisms at the individual level. A theory of structural causation is a theory, but an incomplete one, just as theory at the individual level is incomplete until it tells us what process determined the identities and number of players, why these players value the ends they pursue rationally and which variety of rationality guides their choices, how the institutional arena for the game evolved, where the payoffs come from, why the rules sometimes change in mid-game, and how the power distribution among actors determines the macro outcome. And both microtheories and macrotheories are incomplete until we understand them in their slowly but constantly evolving historical-structural context.

This insistence on bridging levels of analysis is not mere methodological prudery. Empirical questions of great theoretical, even paradigmatic, import, such as whether individuals affect democratization, depend on it. Rational choice theory assumes that they do; Linz and Stepan and O'Donnell, Schmitter, and Whitehead asserted that they do.[14] Yet, despite all the eloquent theorizing that led to "tentative conclusions about uncertain transitions," all the cases covered by *Transitions from Authoritarian Rule* underwent successful transitions that have lasted remarkably long. There are many possible explanations for this genuinely surprising outcome, but one that is plausible enough to require disconfirmation is the idea that these transitions were

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

driven by structural conditions. Even if it is true that elites and groups had choices and made consequential decisions at key moments, their goals, perceptions, and choices may have been decisively shaped by the context in which they were acting. If so, they may have had a lot of proximate influence but very little independent influence after controlling for the context. I do not assert this interpretation as fact but merely suggest that it has some plausibility and theoretical importance; we will never know how seriously to take it until we bridge these levels of analysis with methods that permit testing of complex multivariate hypotheses. This effort would require collecting a lot of new data on smaller units of analysis at shorter time intervals.

Conclusion

Both small- and large-N comparisons have methodological advantages. Small-N comparisons tend to be more faithful to the rich concepts that inspire our theories and tend to be more sensitive to the complex and conditional causal relationships and intertwined levels of analysis that most closely approximate our intuitive understanding of how the political world really works. But no degree of methodological refinement can rigorously justify generalizing the conclusions from a study of a few cases; for such generalization, large-N comparisons are indispensable. Still, the generalizations of large-N comparisons will produce only disappointingly thin tests of theory until they incorporate the conceptual and theoretical thickness of small-N studies. The most practical solution is to combine the advantages of both approaches.

NOTES

I am grateful to David Collier, Ruth Berins Collier, and three anonymous reviewers for their thoughtful and constructive suggestions.

1. Robert H. Bates, "Letter from the President: Area Studies and the Discipline," *APSA-CP: Newsletter of the APSA Organized Section in Comparative Politics*, 7 (Winter 1997), 1–2; Barbara Geddes, "Paradigms and Sandcastles: Research Design in Comparative Politics," *APSA-CP: Newsletter of the APSA Organized Section in Comparative Politics*, 8 (Winter 1997), 18–20.

2. Karl R. Popper, *The Logic of Scientific Discovery* (New York: Harper and Row, 1968). Of course, we also need to demonstrate that our hypotheses fit the evidence. Though not always easy, this demonstration 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. Some scholars therefore 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, that is, logical deduction of the implications of simplistic assumptions. I consider this stance an overreaction to the danger of chance associations. The counterintuitiveness of a

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms

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, replicating with different indicators, and testing other observable implications of the hypothesis.

3. Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes," in John Worrall and Gregory Currie, eds., *The Methodology of Scientific Research Programmes* (Cambridge: Cambridge University Press, 1978), pp. 8–101.

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

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

6. Rational choice is primarily a method for generating theory. Rational choice theorists who test their models have no choice but to use existing empirical research methods and are subject to the same methodological standards as quantitative or small-N researchers. I do not take the naive falsificationalist stand that any single disconfirmation invalidates the entire approach; such invalidation comes only after the accumulation of many disconfirmations of an approach's predictions. However, the accumulation of consistencies or inconsistencies requires much prior testing of specific hypotheses. In this process every hypothesis must be evaluated with respect to at least one competing hypothesis, perhaps drawn from the same approach, perhaps from others.

7. David Collier and Steven Levitsky, "Democracy with Adjectives: Conceptual Innovation in Comparative Research," *World Politics*, 49 (April 1997), 430–51.

8. It is sometimes possible to combine multidimensional components into a single indicator. However, a theory that tells one how to combine them properly is required. In geometry, for example, volume is a single indicator of a multidimensional quality, but it can not be calculated unless one knows the appropriate formula for the shape of the object in question.

9. Adam Przeworski, Michael Alvarez, José Antonio Cheibub, and Fernando Limongi, "What Makes Democracies Endure?," *Journal of Democracy*, 7 (January 1996), 39–55; Giovanni Sartori, *The Theory of Democracy Revisited* (Chatham: Chatham House, 1987).

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

11. As Fearon argues, both large- and small-N tests involve accepting some counterfactual propositions. In large-N regression analyses it is assumed that the explanatory variables are not correlated with any omitted variables. James D. Fearon, "Counterfactuals and Hypothesis Testing in Political Science," *World Politics*, 43 (January 1991), 169–95. Only careful refutation of the plausible alternatives can make this counterfactual plausible. A large-N analysis has the potential to deal with the inevitable counterfactuals more satisfactorily, and we should improve measurement and control so that large-N analysis can better realize its potential.

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

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

14. Juan J. Linz and Alfred Stepan, eds., *The Breakdown of Democratic Regimes* (Baltimore: The Johns Hopkins University Press, 1978); Guillermo O'Donnell, Philippe Schmitter, and Laurence Whitehead, eds., *Transitions from Authoritarian Rule* (Baltimore: The Johns Hopkins University Press, 1978).

This content downloaded from 129.74.250.206 on Tue, 06 Dec 2016 14:53:21 UTC
All use subject to http://about.jstor.org/terms