Letter from the Section President: Pluralism as a Hard Choice

Colin Elman
Syracuse University
celman@maxwell.syr.edu

I want to thank outgoing section President John Gerring for all of his efforts and achievements. John worked tirelessly on behalf of the section. Its overall good health—as reflected for example in the section's size and the range of its activities—is a tribute to his leadership. I also want to take this opportunity to thank Gary Goertz for his outstanding editorship of this newsletter, and the section's 2009 Division Chair, Rudy Sil, for constructing a first rate program.

As I begin my term as President, I would like to take a few lines to reflect on the section's epistemic diversity. Although this range is wholly welcome, it also lends itself to what might be described as a "prolate spheroid" problem. It is natural for people to see a particularly shaped ball as the one used for their own game—Americans naturally assume it is a football, as an Englishman I would suppose it was a rugby ball, and so on.

The common shape we see in qualitative research is close engagement with subjects and cases, and data collection initiated at extremely short range. This approach produces thick, rich, and heterogeneous data. These data are then used to advance a range of analytical goals (depending on the episteme in which the research is located) including interpretations, descriptive generalizations, and causal inferences.

I hope that it is uncontroversial to observe that the section includes qualitative researchers whose work is embedded in and justified by different epistememes, including within-case and small-n, interpretive, and statistical/regression approaches. Accordingly, while we may see people doing what we think of as "qualitative research," the meaning of what they are doing, and the rules they will be following, are different as we move from episteme to episteme.

There are at least three different ways to handle this diversity:

Monism establishes a single episteme as the 'right' way. This claim might be based on a variety of arguments, including an assertion of foundational logic, a declared monopoly on the ability to answer privileged types of questions, or the predicate ability to deconstruct knowledge claims.
Eclecticism proposes that epistemic differences need not matter. It leapfrogs over them to a problem-driven pragmatism—use what works, in whatever combination it works.

Pluralism embraces diversity as a virtue, and seeks to find the limits of association and commensurability between several equally valid epistemes.

The section needs proponents of all three strategies taking part in the discussion. In particular, pluralists are needed to balance the monists' one-sized-fits-all, and the eclectics' rejection of foundations.

Without the check of pluralism, pursuing monism will be an unpleasant and ultimately self-defeating process. Whichever of the approaches emerges victorious from the struggle, the section will be impoverished by the absence of those it managed to exclude. As for eclecticism, it may be liberating for individuals to declare "badges, we don't need no stinking badges." But a strategy of decoupling methodology from epistemology simply delays the inevitable reckoning when lurking incommensurability surfaces.

While I believe that pluralism is a necessary part of the conversation, I acknowledge that it is also a hard choice. It requires learning about other research traditions, and being tolerant of differences. It mandates the sharing of scarce resources. It also means becoming more careful about overly broad claims, whether made in the positive ("this approach is the future of political science") or the negative ("scholars studying a few or single cases are historians not political scientists"). Innocuous in the context of a single epistememe, these types of statements are staggeringly insensitive to other research traditions, and antithetical to a section that strives to be inclusive.

As the incoming President, I want to make the argument that the section's breadth is an opportunity to be embraced. It is not just a problem to be overcome by a retreat to homogeneity, or to be ignored by proceeding as if our differences do not matter.

Symposium: Cautionary Perspectives on Multi-Method Research

Is Multi-Method Research Really "Better"?

Amel Ahmed
University of Massachusetts, Amherst
aahmed@polisci.umass.edu

Rudra Sil
University of Pennsylvania
rudysil@sas.upenn.edu

Recent scholarship in political science attests to the rapid proliferation of approaches engaged in multi-method research (MMR), research that employs two or more methods selected from an array of qualitative, quantitative, and formal methods typically used in the social sciences.1 The general notion that different types of methods can be employed to advance or test a particular theory is not in itself new. Multi-method approaches have long been a feature of social science research, taken up usually out of necessity (e.g. Jick 1979). Where data conducive to one method was not available, scholars would turn to another in order to fill the gap.

What is different about the more recent movement towards MMR is the extent to which the use of multiple methods is undertaken self-consciously by a single scholar in a single work in relation to a single research question, predicated on the assumption that the use of different methods will yield better results in addressing that question. Indeed, for some, MMR has come to represent not a pragmatic response to the complexity of a given problem but an end in itself and even a new universal standard for good scholarship. The method wars in political science have given way to an apparent consensus on the primacy of MMR as a way to achieve better results and overcome the limitations of particular methods.

This trend is most apparent in comparative politics where some of the field's leading figures have called for MMR as a means of overcoming the limitations of single-method research (SMR). According to George and Bennett (2005), the differences between different methodological approaches give them "complementary comparative advantages." Many have gone beyond this, however, to argue that MMR in fact represents a better research strategy, producing greater inference and more reliable findings than SMR (Brady and Collier 2004; Lieberman 2005; Laitin 2002; George and Bennett 2005).

While the increased appreciation of diverse methodological approaches is a welcome trend, there is much to be alarmed about in the recent rush to assert the primacy of MMR. What began as a movement to employ "all means necessary" in the service of problem-driven research is quickly turning into a new dogma that researchers must, or ideally should, incorporate "all means available" to validate their work. Practitioners and advocates of multi-method approaches explicitly argue that, all things being equal, multi-method research is better than single-method research. The goal of this essay is to assess this claim. While "better" can be understood in many different ways, our interest is particularly in the epistemological viability of the claim. Do we in fact learn more by combining different methods? Does MMR actually increase the number or strength of the inferences we can derive in the course of carrying out an individual research project? Can multi-method research actually increase the validity of findings? We are not concerned with the limitations or particular shortcomings evident in works using MMR. We are more interested in working towards an understanding of whether in principle MMR should be privileged over SMR at all times. In other words, is good
MMR always better than good SMR?

Our argument is straightforward: The claim that MMR is inherently better than SMR is built on the faulty premise that one method can offer external validity for the findings of another. Different methods can at best corroborate each other’s findings, but this does not yield a more compelling inference. We do not know more or know better as a result of triangulating different methods because different methods rest upon incommensurable epistemological foundations that even the most heroic attempts at translation cannot overcome. Though combining methods may and often does produce good scholarship, we find that MMR holds the same epistemological status as separate projects addressing the same question, and that SMR is no less likely to produce good scholarship. The “goodness” of scholarship ultimately depends on the care and originality with which research is designed and executed, not on the number of methods that are deployed. Thus, although MMR is certainly valuable for social science research and should be welcomed as a part of a broad repertoire of methods available to scholars, there is no epistemologically sound reason to elevate it above others. Below, we make this case by examining some common forms of MMR and then considering some of the hidden costs associated with prodding individual scholars to adopt a strategy of MMR in the context of a single project.

Statistical Analysis and Varieties of Case-Study Approaches

Despite the great expectations of social scientists, practical applications of MMR have revealed the limitations of trying to use different methods to introduce measures of external validity. Because methods are premised on different epistemological commitments, they tend to employ different types of variables or mechanisms and privilege different dimensions of social reality. They also tend to focus on concepts that can be operationalized within certain boundary conditions and thus may not easily translate to other modes of inquiry premised on different sets of theoretical priors. For these reasons combining methods often results in findings that are incommensurable, frustrating efforts to offer external validity.

Take for example one of the most popular forms of MMR in use today: the combination of statistical and case study analysis. Proponents of such an approach argue that it offers greater analytical leverage, as one method compensates for the limitations of the other. Statistical analysis is limited in its ability to identify a causal mechanism as it tends to focus on causal effects. Case study analysis, on the other hand, is well suited to identifying a causal mechanism, but is has limited generalizability. The combination of the two, it is argued, gives us the best of both worlds. Statistical analysis can be used to identify a general distribution of causal effects whereas the case studies may be employed for the purpose of identifying causal mechanisms and revealing separate links in a causal chain (Lieberman 2005; George and Bennett 2005).

First, it should also be noted that efforts to combine statistical and case-study analysis are only limited to certain kinds of case studies—those designed on the basis of an empiricist epistemology. Empiricism, which emphasizes the temporal priority of positive empirical observations and thus privileges inductive logics, provides a foundation for the probabilistic worldview of statistical analysis and is also consistent with the use of case studies for testing hypotheses or developing hypotheses that can be subject to quantitative tests. But, case studies have a variety of purposes. Some are constructed so as to support a more deductive orientation to theory building, as in the case of "analytic narratives" (Levi 2004) used in conjunction with game theoretic models. Others are designed on the basis of more hermeneutic or phenomenological approaches that stress the interpretation of meanings held by actors within distinctive contexts (Yanow 2006).

But, even with a common empiricist orientation, the claim that combining the two methods offers greater analytical leverage is difficult to sustain on epistemological grounds. This is because in moving from one mode of inquiry to another the basic conceptual categories will necessarily shift. McKeown (2004: 140–146) makes the point that the kind of inference that is privileged in the quantitative worldview cannot be conflated with the broader process of scientific inference which, in the case of a case study, focuses attention on a fundamentally different task: explaining how a set of initial conditions enable particular mechanisms to have particular effects in one or more contexts. Thus, in moving from large-N statistical analysis to case study analysis, the case study will by its very nature introduce variables not present in the statistical analysis (for example, variables that are not quantifiable but whose effects can be observed within a given context). It is precisely the depth of inquiry that enables case study analysis to identify causal mechanisms and reveal their effects in relation to a given outcome. However, it is also this feature that makes the two modes of inquiry incommensurable. They are effectively examining two different sets of variables. Thus in this scenario, the case study cannot be said to either confirm or falsify the finding of the statistical analysis. Though it may offer a plausible story, at an epistemological level, it offers no corrective for the built-in limitations of the statistical analysis. In such combinations, the statistical analysis and case study analysis effectively represent independent (thought perhaps complementary) intellectual exercises. The findings of one cannot be said to validate the other.

Following the reverse sequence does not make this problem any less intractable. For example, if one begins with case study analysis to establish a causal mechanism and then tests the causal relationship with large-N statistical analysis, the latter cannot serve as a source of external validity for the former. As George and Bennett (2005: 138) note, causal mechanisms operate at the ontological level and can be neither conflated with, nor subsumed under, hypothesized causal effects. Statistical analysis can validate the relationship between hypothesized causal effects and generalize it across cases, but it cannot show that the causal mechanism found in the original case study analysis operates in the same manner and produces the same effect across different spatial and temporal contexts. Without case study analysis of each case, one cannot verify that the same causal mechanism is at work in all, let alone approximate the general size of its effect upon a given dependent
variable. Both can provide independent analysis of related research questions, but one does not validate the other.

**Formal Modeling and Empirical Analysis**

Another popular mode of MMR features the combination of formal modeling with some form of empirical analysis. Proponents of such approaches also make the claim that the combination of the two methods produces greater insight, as one can be used to compensate for the limitations of the other. The formal model is thought to provide analytical rigor while the empirical analysis grounds the investigation in some social context. With this sort of combination the different methods tend to have somewhat more defined roles: the formal model is used for deductive theory-building and the empirical analysis is used for the purposes of illustration or theory-testing. However, the two sets of methodological operations involve quite different foundations. Even where the empirical analysis and formal model are both conceived of as broadly positivist enterprises, the former follows from empiricism and the latter from logicism (Shapiro and Wendt 2005). Using either case studies or statistical analysis in conjunction with the construction of a formal model would require completely ignoring the foundational principles on which the latter is built. The empirical analysis is also constructed specifically to test deductive models rather than to generate alternative causal stories since the causal structure and explanatory logic of the model depends more on axiomatic principles and internal consistency rather than on inferences from observed regularities.

In the case of “analytic narratives” (Bates et al. 1998), for example, the case-specific narratives certainly provide context, but the causal story does not emerge from this context, and the interpretations of the contexts are not evaluated against the strength of other interpretations of the same contexts. Moreover, since “theory means formal theory” (Bates et al. 1998: 3), there is the question of how the deductivist logic of formal theorizing can be meaningfully combined with the interpretive logic informing the construction of a context-bound narrative. While it is indeed admirable that scholars are able to demonstrate their use of both extensive-form game theory and case-specific research (sometimes involving fieldwork or archival research), the core causal logic of the explanation is given by the assumptions and logics built into a particular game-theoretic model rather than a balanced process of moving between theory and data (Sil 2000: 375). To be sure, empirical analysis can reveal flaws in an existing model and potentially inform a new model, but this can be done in differently designed SMR projects motivated by different objectives as part of our collective efforts to further knowledge. There is no inherent value to insisting that a single study incorporate both a formal model, premised on a logical positivist worldview and partial to deductive theorizing, and a case-specific narrative developed through interpretive methods designed to generate a deeper understanding of a given context rather than a more general explanation. Thus, while both game theory and case study serve important roles, those roles are distinct, each dictated by the nature of the methods combined, and each producing distinct, fundamentally incommensurable research products.

In the case of the marriage between statistical and formal approaches (e.g. Goldthorpe 1997), too, the use of one method to validate the findings of another is highly problematic. Although statistical analysis and formal modeling both proceed from a broadly positivist foundation, as noted above, the inductively oriented empiricism undergirding the former is fundamentally at odds with the deductively oriented logicism of the latter. The empirical analysis may corroborate the findings of the formal model by capturing the expected distribution of outcomes, but it cannot capture the effects of cognitive mechanisms that are predicated on the assumption of instrumental rationality.

Statistical analysis can determine a general pattern corroborating the findings of a formal model, and case study analysis may offer an illustration of a particular dynamic, but neither mode of inductive inquiry can offer external validation of a model deduced from a priori axioms. Again the problem of epistemological incommensurability presents an obstacle in the quest for external validation. Though considering empirical evidence maybe a useful heuristic for formal modeling, the juxtaposition of the two does not strengthen the inference we can draw from the former or the confidence we can have in a given model.

**The Hidden Costs of MMR**

The argument up to this point is not that there is no value in examining a substantive issue through the application of different methods. It is that, in the common forms of MMR considered above, what one can learn by juxtaposing two or more methods within a single research product is not fundamentally different from what we would learn from separate studies using different kinds of SMR to address the same question. This is not to say that the findings of a single research product featuring MMR are not useful. But, epistemologically speaking, there is no intrinsic gain from insisting on always triangulating different types of methods within a single approach rather than encouraging scholars interested in different methods to use those methods to explore the same substantive question as part of a larger collective effort to generate insights into the question. The combination of methods in a single study does not resolve the problem of epistemological incommensurability, and thus cannot eliminate the tradeoffs built into each of the methods employed at various stages of a multi-method project. Thus it cannot be said that we know more or know better when multiple methods are deployed.

This critique may ring hollow to some. Even if MMR is simply a juxtaposition of different intellectual projects within the same work, it might be argued, is this not in itself an advance over SMR? More significantly, is it not more efficient to have a single scholar generate findings using different methods when investigating the same sorts of substantive problem? Our conservative answer to both questions is: perhaps in some rare instances, but generally not.

Here, the main argument concerns the ways in which research is organized in the discipline writ large and the ways in
which individual scholars conceive of their roles and contributions within the discipline. One must consider the amount of time and energy that is involved in MMR in terms of the methodological training, the fluidity in applying each of the methods, and the total time spent conducting research. Here, three possibilities exist. First, a scholar may end up taking the same amount of time to produce a single product using MMR as it would have taken to apply different methods in different research products. Second, time pressures related to granting deadlines, promotions, or other professional considerations may force researchers to spread a more limited amount of effort over different pieces of the research product. This may ultimately hurt the quality of the scholarship, producing thin case studies, shoddy datasets, and unsophisticated models hastily put together to round out a multi-method project. Third, it may actually end up taking more time as a single scholar shifts gears from one phase of a project featuring one method to a different phase using another method, especially if the scholar requires more "retraining" or "retooling" to effectively apply different methods. In all three scenarios, either some component of the research product will suffer, or a heavy burden will be placed on individual scholars.

Besides the costs to the individual researcher, there are considerable costs to the discipline as well. As researchers become more and more diversified in their methodological skills, they will likely become less and less specialized, spending less time on the approach they are most skilled at or most passionate about. This will have the inevitable effect of diluting the pool of expertise in the field and decreasing our collective efficiency as each of us feels compelled to maintain proficiency in the application of quite different types of methodological tools. Where there is a pool of labor available, it is not at all clear what is gained by making all of the members of that pool make the same kind of investment in gaining the same array of skills. If anything, the gains to efficiency from more specialized training and from iterated applications of the same method are lost. While specialization can certainly be taken too far, there is no reason to think that the entire discipline will gain by having all its members trained to do multi-method research.

Finally, there is a danger that the move towards MMR will result in decreased scrutiny of the core assumptions and epistemic commitments underlying the methods we choose to employ. Because MMR holds out the hope that by the very act of combining methods one can somehow overcome their individual limitations, it also holds the danger that scholars will feel justifiably absolved from questioning the foundations of the multiple methods they employ. This is perhaps the greatest danger of MMR: that we as a community of scholars will lose our critical eye towards methods and, in the process, lose our awareness of the fundamental challenges that have accompanied social scientific inquiry from its very inception.

Conclusion

The notion that MMR is always better than SMR assumes that if MMR is done well, one method can be used to validate the other, and a single project can generate more robust results or more compelling findings. The reality of social science research, however, is that all methods have limitations. This is no less true when they are deployed simultaneously in order to investigate a certain research problem. The problem of epistemological incommensurability that has long plagued exchanges across different research traditions is not being resolved by MMR; it is simply being transferred from the level of the discipline or subfield to the level of the individual scholar and research project. In this context, combining methods may produce complementary results, which may be valuable from a pragmatic point of view in generating support for a particular proposition. However, from an epistemological point of view, such combinations have the same structure and value as separate studies addressing the same research question. Putting the burden of producing these separate studies on a single scholar may not produce any gains to efficiency and may, in fact, come with some significant costs both for the scholar and for the discipline as a whole. In light of these concerns, MMR should be thought of not as the new gold-standard in research but as part of a diverse repertoire of methodological approaches that may be useful to some scholars depending on the nature of the research question—and on the preferred skill-sets and intellectual passions of the individual scholar.

Notes

1 This essay is part of a longer article that is currently under preparation. It also builds on Amel Ahmed and Rudra Sil, "The Logic(s) of Inquiry: Reconsidering Multi-Method Approaches." Working Paper No. 16 of the Committee on Concepts and Methods, International Political Science Association (November 2008).

2 King, Keohane, and Verba (1994, 85–87, 225–227) view causal effects as logically prior to and more reliable than unobservable mechanisms. For KKV, the value of mechanisms is limited to their ability to generate new observations that may influence the level of confidence in causal inferences. David Waldner (2007: 154) interprets this position as flawed in that it fails to recognize the epistemological function of mechanisms, reducing them to mere "servants of inferences."

3 Moreover, as explained by Rohlfling (2008) in the cases of "nested analysis," the introduction of case studies introduces bias potentially exaggerating findings.

References


findings point in the same direction—statistical significance and coefficient signs match the outcome of a case study—does not make them any more likely to be true, since the concepts applied in one methodological component are not equivalent to those applied in the other. It is impossible for qualitative and quantitative methods to say the same thing because they are talking about different things. An alternative schema, though, is possible based on seeing the component of an MMR design as representing two distinct cultures of inquiry that are complementary, rather than corroborating.

**Conceptual Stretching and Causal Analysis**

The topic of conceptual stretching was first broached as a warning against the proliferation of quantitative, statistical methods and the incessant drive to substitute quantifiable variables in place of qualitative categories. Conceptual stretching occurs, according to Sartori, when “denotation is extended by obfuscating the connotation” (Sartori 1970: 1041). The term itself is used in the context of the ladder of abstraction, in which concepts are mapped along two inversely-related dimensions: the intension (connotation), the systematic and explicit definition of the characteristics of the concepts, and the extension (denotation), the range of cases which can be categorized as meeting the conceptual definition. Increasing a concept’s extension by incorporating more cases under its rubric leads to stripping away some of its necessary intension, the specificity of characteristic involved in the concept. While ascending or descending the ladder is critical for theory building, Sartori deplores attempts to expand a concept’s extension without acknowledging concomitant diminishment in intension. Collier points out some exceptions to the law of inverse relation, but maintains the fundamental argument that in travelling to cover new cases, conceptual definitions can suffer unacknowledged distortion (Collier and Mahon 1993; Adcock and Collier 2001). Such distortion makes a concept’s terms, definitions, and referents inconsistent, violating a crucial criterion in evaluating social scientific work (Gerring 2009: 112).

To understand the nature of conceptual stretching in MMR better, a closer comparison of qualitative and quantitative approaches to conceptualization is necessary. Coppendge (1999), an early and eloquent advocate of MMR, describes qualitative concepts as “thick,” having complex definitions developed iteratively through examination of a small realm of cases. In contrast, quantitative concepts are “thin,” with relatively simple conceptual definitions. Conceptual thickness/thinness is inversely related to narrowness/breadth of extension. Because of their definitional intricacy and high intension, qualitative concepts are designed to apply to only a small number of cases. The proliferation of vocabulary about various democratic and authoritarian regime types and subtypes is exemplary of the type of highly descriptive conceptual categorization used in qualitative analysis. Quantitative scholars, by comparison, rely on datasets like Freedom House or POLITY, which reveal only that two countries are equally democratic (or undemocratic) and have no substantive meaning to the distance between intervals (Munck and Verkuilen

---

**The Challenge of Conceptual Stretching in Multi-Method Research**

Ariel I. Abram  
University of Oklahoma  
ariel@ou.edu

Multi-method research (MMR) has gained enthusiastic support among political scientists in recent years. Much of the impetus for MMR has been based on the seemingly intuitive logic of convergent triangulation: two tests are better than one, since a hypothesis that had survived a series of tests with different methods would be regarded as more valid than a hypothesis tested only a single method. In their seminal Designing Social Inquiry, King, Keohane, and Verba (1994) argue that combining qualitative and quantitative methods is useful because it increases the amount of data used to test a specific theory or hypothesis. While critical of specific prescriptions in KKV, Brady and Collier (2004) and Gerring (2007, 2009) reiterate the mantra of epistemological monism, shared standards, and logical consistency between qualitative and quantitative methods.

This paper, though, warns that what Sartori (1970) calls the “stretching” or “straining” of concepts between qualitative and quantitative domains has potentially damning implications on MMR. Simply because qualitative and quantitative
On the other hand, the simpler definitions and low intensity involved in quantitative concepts are amenable to incorporating much wider universe of cases for interrogation via statistical analysis. Thus, concepts like development as measured in per capita GDP can be applied widely and easily analyzed in terms of scalar relationships of cases (Ragin 2000).

Fundamentally, the difference between qualitative and quantitative approaches to measurement is ontological. Since measurement is about relationships of objects to one another, a change in measurement will lead to a different set of relationships (Franzosi 2004: 281). Qualitative measurement necessarily involves the use of categorizing cases using a specific nominal definition. For such a schema to be meaningful, membership must be absolute and each case must be equivalent in the same set of crucial characteristics. The goal is to eliminate measurement error by devising a comprehensive conceptual framework in a clearly delineated case domain. By contrast, quantitative measurement involves scoring cases so that every case is related to the other on an interval scale (Mahoney 2000; Goertz 2008).

To some, the danger of measurement error and problems of causal inference which can arise when quantitative indexes do not relate closely to a given concept is more apparent than real. After all, qualitative concepts can be represented in a quantitative equation using categorical (i.e., “dummy”) variables. Still, there are several problems with this solution: the more complex a conceptual definition becomes, the more difficult it becomes to summarize using quantitative measurement across a large sample and the more tenuous the crucial assumption of unit homogeneity becomes (Gerring 2007: 52). In a statistical sense, the use of dummy variables represents a significant sacrifice in analytical power and cannot be arrayed in a meaningful interval format (Taagepara 2008: 58, 299–231). Moreover, ‘translation’ between qualitative and quantitative data is not transitive: while qualitative data might travel into a quantitative domain, the opposite translation is difficult, if not impossible (Creswell and Clark 2006: 128).

Moving between qualitative to quantitative concepts, then, is bound to cause some distortion or mismatch in the intension and extension in qualitative and quantitative concepts. While concepts may share a label or term, they have different characteristics and refer to different sets of cases with varying attributes. This is depicted in Figure 1 below where the horizontal axis lists four characteristics necessary for falling under the concepts rubric (A, B, C, and D) and the vertical axis the range of empirical cases that fall under it.

In the figure, conceptual stretching is most obviously manifest at the level of extension: numerous cases that would be counted as “in” in a quantitative setting but “out” in a qualitative setting. But a more troubling slippage occurs at the intensive dimension, where characteristics that are crucial in the qualitative definition of a concept are absent in the quantitative one. As Goertz notes, causal hypotheses are embedded within a concept’s connotation. In the qualitative setting, specific causal hypothesis are dependent on cases having attributes A, B, C, and D. These attributes explicate the mechanisms by which causes produce their effects (Goertz 2006). Cases that do not share these attributes, such as those in quantitative domain that share only A and B, cannot be considered candidates for at least some of the mechanisms hypothesized in the qualitative domain.

This insight is crucial whether a variable appears as explanandum or explanan. For example, consider Sen’s critique of quantitative, income-based headcount conceptualizations of poverty and argument for thickening the concept by considering social exclusion and relative inequality (Sen 1987). This revision changes who is counted as poor and therefore forces an explanation of poverty that accounts for a different universe of incidents. If the category of the “impoverished” includes victims of social exclusion who might still earn above

---

**Figure 1: Overlay of Thick and Thin Conceptualization**

![Figure 1: Overlay of Thick and Thin Conceptualization](image-url)
a given standard of living, then a range of solely economic policy prescriptions are inadequate to effectuate change. Alternatively, the addition of these social dimensions to the concept of poverty also has implications for studies of the poverty’s effects on outcomes such as democratization, terrorism, or civil war, where the definition of the dependent variable is often belabored but the independent variable left merely as dichotomous. Conceptualization is always implicitly a process of defining a scope condition, delineating a set of cases with common attributes which are hypothesized to have the same etiological properties. If one or more of these attributes is present in the qualitative conceptualization but absent in the quantitative conceptualization, this circumscribes the range of properties to which causation can be attributed (Lucas 2003; Goertz 2006).

Conceptual Stretching in Practice

Conceptual stretching is a potential challenge in all social scientific work, but it takes particular forms in MMR. One of the most obvious manifestations is alterations between the use of nominal and ordinal modes to represent same variable. For instance, it is logically problematic to claim that five cases are in the set of democracies and then to claim that one of these is “more democratic” than the others. In some cases this is the result of imprecise or sloppy language. But such inconsistency obfuscates what precisely the key variables are meant to account for and the relationship of the cases to one another (Mahoney 2000: 408).

A more subtle form of stretching can be seen in the difficulties of establishing qualitative and quantitative case domains, since case selection is intimately tied to conceptualization (Collier and Mahoney 1996). Evan Lieberman’s Race and Regionalism in the Politics of Taxation in Brazil and South Africa (2003) exemplifies a methodologically astute and sophisticated work that nonetheless exhibits unacknowledged conceptual stretching between qualitative and quantitative renderings of the critical variable. Lieberman’s central contention is that

[the specification of group rights in the form of official state documents and policies provides a stronger set of incentives for political entrepreneurs to make claims based on such identities... Federalism, for example, tends to give important political salience to regional identities, and official racial exclusion tends to give much more salience to racial identities... (Lieberman 2003: 14).

Privileging certain types of identity over another, which Lieberman calls the concept of national political community (NPC), constrains the ability of political entrepreneurs to demand the creation of progressive taxation systems.

Lieberman’s qualitative, historical investigation of Brazil and South Africa comes to a countereintuitive but persuasive conclusion: South Africa’s 1909 Constitution specified white supremacy while denying recognition to regional differences. By establishing whites as a legal category, this cornerstone document encouraged the white economic elite to cooperate with the state in establishing a social safety net system that raised the living standards of their poorer co-ethnics. After apartheid’s downfall this redistributive system was opened to all races, leading to greater economic equity overall. In contrast, Brazil’s 1891 federalist constitution privileged claims based on regional equity, but was explicitly inclusive on racial grounds. Faced with the prospect of racial equality, Brazil’s white economic elite opposed the state’s effort to develop a redistributive system that would benefit a mainly black underclass.

The stretching of NPC becomes apparent as Lieberman attempts to test whether similar racial and regional definitions of NPC the same effect in other cases. Examining constitutions and other legal documents from over one hundred cases, he converts the data into a series of dummy variables. The statistical results show a correlation consistent with the small-N study, but the regression model does not capture important dimensions of the qualitative findings. Specifically, the historical narrative focuses on the interplay of racial and regional identity, but in the quantitative analysis these variables are not interacted. More importantly, the qualitatively-derived elaboration of NPC proves too narrow to incorporate the majority of empirically relevant cases. The initial, thick-rom of NPC is limited to two dimensions and four different modes, as shown in Table 1. It is the concept’s thickness that allows Lieberman to make specific hypothesis about the identity bias for collective action.

Table 1: Distribution of Cases in Lieberman’s Qualitative Analysis (2003: 79)

<table>
<thead>
<tr>
<th>Race Exclusionary</th>
<th>Race Inclusionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>South Africa</td>
</tr>
<tr>
<td>Brazil</td>
<td>Brazil</td>
</tr>
</tbody>
</table>

Table 2: Distribution of Cases in Lieberman’s Quantitative Analysis (2003: 242)

<table>
<thead>
<tr>
<th></th>
<th>Federal</th>
<th>Unitary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragmented</td>
<td>Race Exclusionary</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Race Inclusionary</td>
<td>2</td>
</tr>
<tr>
<td>Not Fragmented</td>
<td>7</td>
<td>43</td>
</tr>
</tbody>
</table>

But in the larger dataset of sixty-nine countries, forty-three cases in the population (62%) have neither relevant racial or regional cleavage that are crucial to the initial definition of NPC and have to be incorporated under residual category of “non-fragmented” states, as shown in Table 2. In the absence of significant racial or regional cleavages, Lieberman’s qualitatively-derived theory about collective action is mute and the mechanisms he posits for explaining collective action cannot account for the majority of the cases in the expanded conceptual domain.
Complementarity Triangulation as an Alternative Logic of MMR

Recent proponents of MMR focus assume the easy commensurability of concepts and focus mainly on whether or not quantitative or qualitative results confirm one another. If the findings of one method do not “fit” the initial model, then its hypotheses are rejected and specification amended (Liebman 2005; Fearon and Laitin 2008). A case in point is the Political Instability Taskforce, which “having at least two independent approaches [one qualitative, one quantitative] to assessing instability, if they point in the same direction, greatly increases the confidence of predictions” (Goldstone 2008: 7). While Rohlfing (2008) cautions that this approach is susceptible to error because qualitative and quantitative technique might have compounding rather than correcting biases, this essentially adopts the monist assumption of comparability between methods. The problem of conceptual stretching—the mismatch between concepts and variables using in different settings—is more fundamental because it questions the very comparability and compatibility of qualitative and quantitative methods. If what is categorized in a qualitative domain as instability is not equivalent to what is scored as such in a quantitative dataset, agreement can be dismissed as mere felicitious coincidence.

An alternative logic to MMR, however, seeks to leverage the complementary aspect of qualitative and quantitative methods. Such an approach assumes that qualitative and quantitative research methods construct different objects, do not measure the same objects, or focus on different attributes and properties of the object, which in sum make mutual validation impossible. Rather,

"the linkage of quantitative and qualitative methods does not produce one unique picture of reality. Instead, the different research results would have to be combined in a sort of jigsaw puzzle to produce an adequate image of reality. (Erzberger and Prein 1997: 144)

The fitting-together of this puzzle depends crucially on the differing theoretical underpinnings of the use of each method, particularly their ability to offer explanations at different levels of analysis. Each method offers a unique capacity for explanation. Contrary to the monists, Mahoney and Goertz (2006) argue that qualitative and quantitative methods represent distinct cultures of inquiry, making inherent some friction between the two. Specifically, the use of nominal measurements in qualitative methods is tied to an approach which focuses on the “causes of effects” in individual cases. In contrast, the use of interval measurements in quantitative methods is tied to an approach focused on estimating the average effect of independent causes. These traditions each offer their own standards by which to evaluate causal claims, but these standards are in a sense complementary:

an explanation of an outcome in one or a small number of cases leads one to wonder if the same factors are at work when a broader understanding of scope is adopted, stimulating a larger-N analysis in which the goal is to explain particular cases and more to estimate average effects. Likewise, when the statistical results about the effects of causes are reported, it seems natural to ask if these results make sense in terms of the history of individual cases, one wishes to try to locate the effects in specific cases (231).

An example of this type of cross-cultural dialogue comes from multi-volume collaborative study of civil wars by Collier and Sambanis (2005). This work uses case studies by individual country experts to explore a statistical model that analyzes the impact of opportunities to seize power (“greed”) or capitalize on thwarted political ambitions (“grievances”) on civil war onset. In reviewing the contribution of the qualitative component, Sambanis (2004) notes that several of case studies identified country-year intervals as periods of civil wars (“1” in the dichotomous dependent variable) which were coded as non-war (“0”) in the dataset. This is not simply an example of measurement error in need of correction. Rather, it goes fundamentally to the connotation inherent to the concept of civil war itself, which has ramifications for the logit/probit results obtained in the statistical analysis. Similar conceptual revisions were made of independent variables. Statistical evidence shows that states with lower levels of education have a greater likelihood of suffering from civil war, since citizens with little education are more likely to be able to “do well” by war by engaging in criminal behavior during civil war. But some countries with relatively well-school citizens, like Cyprus, Yugoslavia, Georgia, Russia, and Lebanon, experienced civil war, while Saudi Arabia, with a low education rate, has not. Qualitative studies suggest that it is not just the extent of schooling but the type of education which can influence the propensity for war. A curriculum rife with ethnic chauvinism, for instance, can also reinforce the motive for violence. Combining a conceptualization of education as consisting of both qualitative categorical and quantitative interval components opens up the possibility of new types of explanations for the apparent linkage of antecedent and outcomes.

The power of such complementary triangulation comes not in testing and rejecting specific hypothesis, but in refining concepts and thereby the range mechanisms that account for the outcome of interest. Areas of congruence and incongruence between qualitative and quantitative concepts must be investigated inductively. Dunning’s (2007) advice that a researcher code a handful of cases from a pre-packaged dataset manually to ensure familiarity with variable features and underlying conceptual definitions is doubly important in MMR. The most relevant cases for the purpose of conceptual clarification are the most liminal at the independent or dependent variables. Determining the disposition of cross-over cases—whether (or not) they can be considered democracies or to have experienced civil war—is crucial because these cases highlight precisely the criterion of inclusion and exclusion within a categorical set (Goertz 2008; Ragin 2000). The goal is not to amend qualitative and quantitative conceptual definitions until they are equivalent, but to identify their divergences.

Explicating empirical heterogeneity and conceptual incongruity is vital to capturing the multiple and manifest pathways
connecting structural antecedents and outcomes. Like all uses of narrative techniques to trace causal mechanisms, MMR must establish its truth claims by testing alternative explanations within a single case (George and Bennett 2005). But identifying a single mechanistic process is rarely enough to account for the full range of correlation observed in the quantitative analysis. In the example from Lieberman above, the problem is not that the mechanisms specified by qualitative methods are incompatible with the statistical pattern. Rather, the problem is that the mechanisms are logically precluded from functioning in many of the cases within the large-N population. Put another way, a significant portion of the $R^2$ observed in the quantitative portion must have been due to mechanisms other than the ones identified in the study. More attention to the exact parameters of stretching between qualitative and quantitative concepts allows the researcher to look for these alternative mechanisms.

Since inferences about mechanisms are derived from the properties of the concept used in the qualitative analysis, the power of MMR is constrained by the representativeness of the qualitative sample. Qualitative concepts and methods are the fulcrum, highlighting heterogeneity among case units and locating multiple mechanistic pathways connecting initial conditions to outcomes. The success of complementary MMR hinges on exploring the overlap between large and small-N conceptualization. Thick and thin concepts should not be rendered equivalent by assumption or theoretical deduction. Rather, inductive techniques must be used to demonstrate the contours and boundaries of their shared analytic space.

Notes

1 For more on the relationship between intention and extension, see Goertz (2006: Ch. 3).
2 See also, Brady (2003).

References

Ontology, Epistemology, and Multi-Methods

Abhishek Chatterjee
University of Virginia
ac7y@virginia.edu

Enthusiasm for multi-methods research can possibly be ascribed to the prima facie promise it holds for moving beyond, if not resolving, seemingly intractable debates on the relative merits of "qualitative" (historical, interpretive, etc.) versus "quantitative" (i.e. inferential statistical) research methods. The justification of multi-methods rests on the claim that combining a few case studies with a larger inferential—and not descriptive—statistical study manages to capture the strengths of both insofar as the discovery of causal relations is concerned. This in turn lends greater confidence that the relationships being asserted are indeed causal. The specific argument is that since inferential statistics allows for generalization (while case studies normally do not), and case studies are better at tracing what are called "causal mechanisms," combining the two affords us the best of both worlds.

The trouble with this is that scholars seeking to justify multi-methods seem to assume that the question of what constitutes "cause" or "causal mechanism" is unproblematic, and the problem is limited to that of making causal claims. The problem, in this view, is solely epistemological. Epistemologies however do not exist in vacuo; they are both supported by and in turn support ontologies (or metaphysics), which can roughly be defined as presuppositions or innate conceptions about the nature of the world. An insufficient appreciation of this leads to mutually contradictory arguments in favor of multi-method research designs; arguments, which on reflection could not possibly support such designs. Arguments conceding the usual weaknesses of case studies—but nonetheless attempting to justify them—imply a metaphysics that makes it impossible to portray case studies as either necessary or sufficient in causal analysis, which in turn also precludes any justification of multi-method research. In other words, some fundamental concessions—implicitly based on a specific ontology—negate almost all subsequent justifications that could be made in favor of case studies, and by extension, multi-method designs.

The causal ontology often accepted in pointing out the deficiencies of case studies—implicitly or explicitly—is "reductionist" and "regularist," i.e. one which respectively defines causes in terms of non-causal relations and states of affair and affirms that such non-causal relations are regularities in nature. The origins of this metaphysical view can be traced to David Hume (1999 [1748]: 136)—hence often referred to as "Humean." The particular conception of what it means to make 'causal generalizations' is a logical implication of this ontology of causality. Moreover the idea of 'generalization' cannot be separated from the definition of causality here; in other words to say that something is caused by something else is also to generalize in a certain way, namely, by referring to regularities. Though inferential statistics finds sufficient justification in (and in turn sufficiently justifies) this ontology of causality, explanations based on case studies are not consistent with it. Case studies and inferential statistics cannot logically mix if the definition of causality is reductionist and regularist. This also applies to arguments claiming that case studies illuminate causal mechanisms, since the only definition of "mechanism" that is consistent with this ontology is one that sees them as concatenation of variables that occur with some regularity, something that case studies are not equipped to handle. Multi-methods using case studies can therefore never be justified under this metaphysical view.

Yet (1) referring to regularities is not the only way to generalize, (2) causes do not necessarily have to contain generalizations, and (3) it may not be possible to reduce causes to something more basic. In each of these three cases, one can find sufficient justification for case studies (and also independently for inferential statistics), but the usual arguments for combining the two run into logical difficulties. This is because the usual justification for multi-method designs is in fact a confusion of distinct metaphysical views about the nature of causation that are not necessarily complementary. How does one know that the mechanism connecting a cause with an effect in a particular case study is the same mechanism connecting causes to effects in all the other cases? What part of the study does the causal work, the case studies or the statistical analysis? If it is the case study then the statistical analysis should not convince us, and if it is the statistical analysis then the case study should not convince us. This epistemological dilemma arises because the problem is not merely methodological; it involves our fundamental, and most often implicit, metaphysical assumptions about the nature of the world. Let us examine these issues in turn.

That small-N is not merely an epistemological problem becomes evident when we ask under what definition of "cause" should small-N be a problem for establishing causal relations. The answer has to do with statistical theory and the Humean conception of causation that sufficiently—though not necessarily—justifies it. To understand this, let us consider the epistemological and methodological implications of this conception. In other words, given a Humean view, how would one go about discovering causal relations? Now very briefly, Humean definitions come in both deterministic and stochastic versions. Causes precede their effects, and are either necessary, sufficient, or both necessary and sufficient conditions (in the deterministic versions), or increase the conditional probability of their effects (in probabilistic versions). In both cases, every singular causal statement must be an instance of one or more general causal laws. The singular phenomenon itself need not be repeated as long as the unique phenomenon can be shown to be the result of a combination of laws that recur in other singular phenomena. Epistemologically therefore, the singular phenomenon cannot play a role in the establishment of a causal relationship since it is itself dependent on preexisting regularities that have already been established. Both the deductive nomological (D-N) scheme of explanation, proposed most clearly
by Hempel and Oppenheim (1948), and Hempel's (1942) inductive-statistical (I-S) scheme follows directly from such conceptions of causation.

The epistemological problem is that of discovering regularities when many laws are instantiated simultaneously. Under ideal conditions experimentation would be the first best method (this obviously is not unique to Humean views; experimentation as a method is consistent with almost all ontologies of causation, but interpretations of experiments would differ depending on the definition of causality). One way to overcome the problem of simultaneous instantiation would be to isolate individual causes and observe their effects repeatedly to establish lawlike regularities. When we move from the experimental sciences to the social or non-experimental sciences, the goal remains the same, i.e. the discovery of regularities, but this time they have to be detected from purely observational data. This is where statistical models come in. Such models try to approximate the experimental situations described above. These models assume that the data being generated are akin to the result of a series of independent experiments or observations generated from mutually independent processes where nature manipulates the independent or explanatory variable under different background conditions or controls (again, it is also possible to give other interpretations to inferential statistics). Inferential statistics is also consistent with the definition of causes as generalizations; that is, the “regularity” part of the definition, or alternatively the definition of causes as “types.” The latter is obviously because insofar as it informs one of average effects, generalization (over a particular population) is built into the interpretation of inferential statistics.

The link between a reductionist and regularist metaphysics on one hand and inferential statistical methods on the other should be clearer now. The impossibility of fitting case studies into this framework should also be evident. Indeed, some prior discussions in political science have clearly recognized this. For example, Sartori defended comparative case studies as a third-best method behind experiments and statistical studies (1994:16). His argument was that though it is true when it comes to drawing causal inferences, comparative case studies are inferior to either experiments or statistical control, the phenomena that most interest certain political scientists do not occur enough times to lend themselves to statistical studies. The problem with this defense is that the acceptance of the logic of statistical inference entails that a few cases cannot or should not lead us to believe that a cause exists. This is the crux of Lieberson’s (1991) argument against drawing causal conclusions from a few comparative cases (also see Sekhon, 2004). Using the example of automobile accidents, Lieberson shows how fragile our conclusions can be as to the causes of accidents if we rely on only a few cases, assuming that knowledge of causes entails knowledge of regularities. The most logical conclusion in this instance would be to state that given the paucity of cases one cannot say anything about the presence or absence of causes.

Lieberson’s critique applies equally to solutions to the problem that urge us to somehow increase the number of cases by, among other things, performing “within case analyses” by looking at multiple implications or consequences of a particular theory or causal statement within the same case (Campbell 1975: 184–189). But if we assume that regularities are most basic and knowledge of causes entails knowledge of regularities, it is difficult to count multiple implications as an augmentation of the number of cases. For at a given level of analysis, each implication of any causal statement must be considered separately. It is for this reason that statistical models require each observation to be independent. And multiple implications of the same causal statement or theory can never be considered independent from each other. There is a rebuttal to Lieberson’s argument, but as we shall shortly see, it makes sense only within decidedly non-Humean ontology of causation. Within the Humean ontology, Lieberson’s position is very convincing indeed.

Again, early discussions seemed to have conceded this. Still case studies were defended variously as “a first stage of research, in which hypotheses are carefully formulated,” (Lijphart 1971: 685), or as explanations of particular cases for their own sake in light of theory, as in Verba’s “disciplined configurative approach,” (1967: 114–115) among others. In such an approach the researcher seeks to explain the event with the help of established regularities and general causal statements. It is important to note here that though disciplined configurative explanations rest on general laws, the explanation itself does very little to strengthen or weaken the validity of the said laws (Lijphart 1971: 692). Yet these concessions are sometimes accompanied with arguments that cannot easily be reconciled with the former. Thus, for example, Lijphart’s subsequent assertion that such studies can be considered “crucial experiments” if values on the variables are extreme is difficult to reconcile with his statement quoted above. Why should extreme values on variables in one case cause us to reexamine our prior theory, especially since the latter could be based on a large number of cases? The same applies for ‘deviant case’ analyses. As Meckown (1999) has also observed in a slightly different context, a single additional case can never, by this logic, lead us to weaken an original proposition that is, in Lijphart’s own words, “solidly based on a large number of cases” (1971: 692). The problem is that some of Lijphart’s epistemological points about the contributions of case studies make sense only when decoupled from his ontological orientation which seems to underline the bulk of his other points.

Another popular defense of case studies—that such studies are better at handling determinism (Gerring 2004: 347, Munck 1998: 33)—is based on conflation of ontology with epistemology. It is perfectly consistent to have a deterministic and Humean view of causality—indeed, the original Humean view was in fact deterministic and some philosophers have argued that “Hume and indeterminism don’t mix” (Dupre and Cartwright 1988)—and still claim at the epistemological level that statistical inference is the best way to establish this causality. As Laplace observed a long time ago, an (ontologically) deterministic relation can appear to be (epistemologically) stochastic because of ignorance of all relevant laws and initial conditions. It does not matter whether the view of causality deter-
ministic or stochastic as long as causation is reduced, and it is reduced to regularities either deterministic or stochastic. In both situations case studies can never be logically justified as the "first best" method. Similar interpretations can also be given to the use of inferential statistics in political science. As a result, criticisms such as Lieberson's against the use of Mill's methods would still be valid. What we discover from Mill's methods cannot even be considered "cause" in this sense.

The reference above was to deterministic sufficient conditions. But can deterministic necessary conditions justify case studies, as Dion (1998) has argued? Dion's argument protects case studies against the small-N criticism only under extremely restrictive conditions. The argument has more to do with the problems that classical inferential statistics faces in tackling necessary conditions than the inherent strengths of case studies. In fact it could be seen primarily as an advocacy of Bayesian statistics over classical statistics when it comes to necessary conditions.

Since Bayes' rule depends crucially on known probabilities to determine posterior probabilities, its applicability is limited to only certain kinds of systems. To be precise, it is crucial that the mechanism that generates prior probabilities is well-known, and alternative hypotheses have well-known probability outcomes or likelihoods. The prior probability is a source of great debate in both philosophy and statistics (See Sober 2002, for example). It is uncontroversial in cases of systems where there is a clear way of assigning prior probabilities. But it is slightly more controversial in cases where we can't. Then the question becomes what should the prior probabilities be based on? Should they be based on statistical regularities, "common sense," case studies, or subjective opinions? As soon as we ask these questions, we realize that we are back to the square one. Additionally, and more pertinent to the use of such statistics to defend small-N's is the fact that we would have to consider multiple hypotheses with determine likelihoods for effective statistical control; at which point the difference in terms of sample size between classical inferential statistics and Bayesianism begins to disappear. Even this argument, as a result, cannot provide sufficient justification for case studies.

This brings us to the final and most popular set of justifications for both case studies and their incorporation in multi-methods research, namely, that case studies are uniquely suited to discover or enunciate what are called "causal mechanisms," which statistical studies are less able to do. However, "mechanism" is yet undefined. Further, of two possible understandings of the concept (of mechanism), one does not provide any justification for case studies, while the other—while sufficiently justifying case studies—cannot easily support their incorporation in multi-method designs.

If mechanisms are defined as, "in effect, variables that operate in sequence," (Sambanis 2004: 288), or any variation thereof, some of the same criticisms that we started with apply. The difficulty of defending case studies while holding this particular understanding of mechanisms stems from the fact that it implies just another version of the Humean definition extended to intervening variables. It is theoretically possible to multiply the number of steps between cause and effect while remaining steadfastly Humean. Each link or mechanism in a longer chain can be represented by equations that can be construed as statements of regularity and as such the same epistemological concerns that were raised earlier about the confirmation of causal claims with case studies apply here too. Various statistical models such as path models would seem to be the natural recourse. If this is a fair representation of some definitions of causal mechanisms, then again the sufficiency of case studies cannot be defended.

More avenues open up once we abandon either reductionist or regularist (or both) understandings of the concept of "mechanism." But these latter conceptions, though equally supportive of inferential statistics independently, cannot easily accommodate the usual manner of performing multi-method research without running into logical contradictions. "Singularist" definitions of causality hold that singular events and not regularities are more basic. The definition decouples generalizations from the definition of causality (Ducasse 1993; Salmon 1980, 1997). Epistemologically, therefore, one need not look for generalities, and the explanation of a single event or case can count as a causal explanation. Process tracing in case studies receives sufficient metaphysical justification here. But this ontology presents us with a problem. Such reductionist but singularist definitions of causality have difficulty distinguishing spurious causes from "real" causes at the definitional level. One way of overcoming this is to attach counterfactuals to singularist mechanisms. Counterfactuals, however, are very sensitive to contrast spaces. The truth condition of counterfactuals depends on the contrast space of any explanation and therefore causality also becomes context and contrast space dependent in this case. So, for instance, causes of revolutions as opposed to near-revolutions can be very different from causes of revolutions as opposed to non-revolutions, or revolution in country A as opposed to revolution in country B. Generalizations, if any, in this case are "bottom up" and change based on the relevant contrast spaces rather than "top-down" and ostensibly universal. Furthermore since singular events are more basic, there is no expectation that generalizations will necessarily emerge. But if we define contrast spaces with as much generality as possible, for instance in our example above, as all possible cases of near-revolution, and if we call answers to both kinds of questions (the limited and expanded contrasts, respectively) "cause," certain problems recur at the epistemological level in combining methods since there is no presumption that an answer to one question will have any bearing on an answer to the other. Thus though attaching counterfactuals to singular mechanisms suffices to justify case studies, they cannot justifi multi-method research.

For instance Evan Lieberman's (2005) latest attempt to suggest a framework for multi-method research faces this particular problem. He writes that "a nested research design implies that scholars will pose questions in forms such as "What causes social revolutions?," while simultaneously asking questions such as 'What was the cause of social revolution in France?'" (2005: 436) For an answer to both questions to
qualify as "causes" almost necessarily implies a singularist view of causation. Under a regularist view an answer to the second question cannot differ from an answer to the first, and the former has to be at least a subset of the latter. His advice is to start with a large-N analysis and then—in case of robust and satisfactory results—"test" the model with small-N analysis by choosing cases that fall within the average prediction of the large-N model (2005: 437). Why should we expect the small-N cases to be consistent with the large-N predictions? Even if they are, why should we have any confidence that the average prediction of the large-N analysis and case study research point to the same causal relationship? In the absence of robustness Lieberman advises model building and analysis of predictions that fall in the average, and also the outliers (2005: 439–440). The criteria for "robustness" and "satisfaction" must be statistical; it is therefore difficult to see why lack of robustness should motivate case studies. Indeed there are well-known remedies within inferential statistics for such problems as lack of statistical significance or any bias in a model and none of these involve looking at case studies. Note that all the questions raised here do not imply that Lieberman is wrong, but that the argument contains large gaps, owing to insufficient appreciation of the metaphysical implications of methods. Additional arguments have to be supplied to reconcile mixing of the two methods.

Case studies also receive sufficient justification if we abandon a reductionist view of causation or causal mechanisms. This would reverse the order of priority in the relationship between regularities and causes. Instead of regularities being signifiers or definers of causes, prior knowledge of causes would restrict and inform the kinds of inferences one is able to make from statistical relations. This is also an effective rejoinder to Lieberson's criticism of the comparative method. This is part of Nancy Cartwright's argument for considering causal "capacities" as primitive. She contends that it is the arrangement of capacities in certain ways that produce regularities; "nomological machines," or "socio-economic machines" as she calls them, are particular arrangements of capacities that "in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behavior that we represent in our scientific laws." (1999: 50) Capacities, further, cannot be identified by any particular manifestation. They can be compared to qualities such as kindness or tenacity that are carried by human beings. Such qualities are not identified with any one particular behavior; instead they are instantiated in multiple circumstances as different behaviors all of which have in common the fact that they are displays of kindness or tenacity (1999: 51). Socioeconomic machines are essentially fables that illuminate important aspects of how the world works, while capacities can be equated with morals of such fables. The relationship between the fable and the moral is not that of similarity but "that of the general to the more specific...[e]ach particular is a case of the general under which it falls" (1999: 39). This means inter alia that "satisfying the associated concrete description that applies on a particular occasion is what satisfying the abstract description consists of in that occasion" (Ibid). Thus any particular arrangement of capacities is also general, and in turn, every general capacity finds its manifestation only in particular arrangements. Once we understand capacities well enough (as is the case in certain natural sciences) we can further manipulate these capacities and arrange them in different ways to produce different laws. As Cartwright observes, "anything can cause anything else. In fact, it seems...not implausible to think that, with the right kind of nomological machine, almost anything can necessitate anything else" (1999: 72).

The epistemological consequences of this view urge us to treat both (most) large-N statistical studies and case studies as essentially alike in that both can be interpreted as attempting to 'guess' the arrangement of hypothesized capacities in the world. Sometimes when we know about enough capacities and other background conditions "[w]e accept laws on apparently slim experimental bases...[and] the data plus the description of the experimental set-up deductively imply the law to be established" (Cartwright 1999: 93). Case studies, both single and comparative, can therefore be considered similar to fables that substantiate morals. The fables however have to be very carefully constructed with great attention to capacities and their arrangements. They are necessarily concrete, but they are at the same time general. This is precisely why studies like John Gaventa's (1980) of one particular locality in one country are also general. Notice that domain restriction finds its best justification under this ontology. In fact if we follow this logic, restrictions of domains is imperative, since what we are describing are particular nomological machines, the very definitions of which carry the connotation of restriction. This is because as we observed earlier, it is the arrangement of particular capacities in certain orders and under certain conditions that could generate laws. But domain restriction does not mean restricted generalization. The fact that some physical laws are literally true only within the confines of the laboratory does not prevent them from also being general. This answers certain criticisms of case studies based on their domain restriction. Thus to say that domain restriction in case studies necessarily implies limited causal force is to implicitly accept an ontology that cannot justify case studies in the first place.

Even in this case, however, the usual manner of combining case studies with a large-N (inferential) statistical analysis cannot be logically supported because of the reasons pointed out earlier. On the other hand, one way of avoiding the usual contradictions in mixing methods would be to truly "trianulate" within the general framework of a case study. In other words instead of using the usual procedures of picking one case out of any sample, one could try to empirically describe the arrangement of capacities (of course, in the context of prior background knowledge of capacities) of any one case, and then examine the implications of such an arrangement using quantitative evidence. This would work because as Cartwright pointed out, it is the particular arrangement of capacities that produces regularities. But it must be a necessary preliminary to first explain why and how the arrangement of capacities came to be. This kind of suggestion is most relevant to the literature on institutions in political science and sociology, especially the ones based on single cases.
Notes

1 Another possible but independent reason, particularly of interest to those interested in the sociology of knowledge is that multimethod research, especially when used in doctoral theses signals to potential employers competency in both statistical and others kinds of research methods thus satisfying the largest possible coalition of potential employers. To reiterate, this is one possible hypothesis in need of further study, and will not be addressed further in this contribution.

2 Though modern versions are significantly different from what Hume originally may have suggested.

References


about what to test and when to test. The third is the need for clarity about specifying appropriate boundary conditions.

To begin, there is a flaw in any multimethod work that relies on mismatched concepts—which probably includes most multimethod research. Different approaches have a strong tendency to use radically different concepts. Case studies employ thick concepts—complex, nuanced, multifaceted concepts, often even concepts with proper names; and quantitative analysis and formal modeling almost always employ thin concepts. There is a lot of information in any concept of “Silvio Berlusconi” that is not contained in any definition of “prime minister,” much less “head of government,” which are concepts more typically used in quantitative analysis; and these contain more information than “agent,” “formeateur,” “chair,” or whatever the closest corresponding concept would be in a formal model. Silvio Berlusconi is indeed a prime minister and a formeateur of cabinets; but he is not just a prime minister and formeateur.

This raises questions about whether it is safe to take an explanation of Italian outcomes and test it more generally with evidence about Gordon Brown, Stephen Harper, Manmohan Singh, or other prime ministers. Maybe it is, maybe it’s not; it all depends on whether the characteristics of Berlusconi that are relevant for the test are ones that are shared with all the other prime ministers rather than any of his more colorful or scandalous characteristics. This difficult translation applies in the other direction as well. A hypothesis that is deduced to be true of formeateurs under a stylized set of assumptions may not apply to actual prime ministers in general, or to Silvio Berlusconi in particular. Some of the stylized assumptions may not be true for them, and they may possess different motives, powers, or constraints that override the assumptions of the model. The mismatch between the concepts calls the relevance of the test into question.

These are major conceptual leaps that must be made carefully. We have to make sure that concepts that are used in one approach are the functional equivalents of concepts used in another approach if we are to get much benefit from combining the approaches. This means that as we climb the ladder of abstraction, we must leave behind the attributes that are irrelevant and take with us all the attributes that matter for the theory at hand. Unfortunately, knowing which attributes matter is hard. It requires round after round of theorizing and systematic testing.

My hope is that eventually we will develop a set of standardized concepts that can be used safely in any approach—concepts that have been repeatedly shown to be useful because they (a) identify consequential attributes of the political world and (b) have robust and plausible causal relationships with other useful concepts. I am not advocating the creation of a central authority that certifies some concepts and indicators as useful and bans others, however. The kind of standardization I have in mind would emerge naturally from the competitive efforts of many scholars and the judgment of their peers, who will slowly come to agree on the most useful ways to define and measure common concepts in comparative politics.

The second problem is that there is still considerable disagreement about what to test and when to test. From the deductive side, there are some who believe that the current emphasis on testing gets in the way of the use of model-building to provide insights, generate surprising hypotheses, and explicate causal processes. Clarke and Primo, for example, frustrated by the higher status accorded models that generate testable predictions, ask, “Why test predictions from a deductive, and thus truth-preserving, system? What can be learned from such a test? If a prediction is not confirmed, are assumptions already known to be false to blame?” (Clarke and Primo 2007: 741). Their answer is that models serve other valuable purposes—providing insights, organizing empirical generalizations, generating surprising hypotheses, defining causal mechanisms, and forecasting the future—and should not usually be subjected to testing (743).

From the inductive side, there are those who believe that it is not necessary to test conclusions that come from intimate engagement with detailed, rich historical, or case study evidence. There is a feeling in some quarters that if you know your cases really well, it is not necessary to see whether your conclusions hold up with systematic, fresh evidence. This stance is most evident in comparative historical research, in which shuttling back and forth between theory and evidence, which entails amending or supplementing the theory in the light of new evidence, is praised as a methodological virtue even though it blurs the line between theory development and testing (Mahoney 2003).

I object to both attitudes. I do respect the need to develop theories and models, drawing on both deductive and inductive approaches. This is a valuable, indispensable stage of the scientific process. But at some point, theories and models have to be tested systematically with fresh evidence, even if there is a presumption of truth arising from inductive engagement with the facts, and especially if they preserve the “truth” of assumptions that are known to be false. Otherwise, deductive approaches risk developing ever-more complicated models that have no connection to reality, and inductive approaches risk giving up on developing theoretical understanding altogether. I get annoyed when either formal modelers or comparative historians claim to have achieved “insights,” to have made “progress,” or to have “valuable” findings before anything has been subjected to a well-designed systematic test. Making good time on the highway is not progress if you’re driving in the wrong direction. There have to be reality checks along the way. Testing is the equivalent of stopping to check the map or ask for directions from time to time.

Testing does not necessarily mean running regressions. That is certainly one kind of testing—extensive testing—which is appropriate when the goal is to generalize, whether to the globe or to the many observations to be found within a single country. But there is also intensive testing that tests which of several hypotheses is most consistent with many kinds of evidence from a single case. I hope that scholars will increasingly recognize that these are two equally valid, yet quite distinct, modes of testing. Neither do I mean to imply that everyone must test, all the time. Clearly there are benefits to having
some division of labor. As a practical matter, it is important for some research programs to develop without testing for a while. But I do think that no project should go too far before someone does some testing (Green and Shapiro 1994).

I also think that we should make some effort to set off in directions that are likely to be right. That means, as much as possible, starting out with assumptions that are fairly realistic. Of course, no assumption is entirely true. All models simplify reality. That’s what models do. But there is a crucial difference between assumptions that are reasonable simplifications and assumptions that are patently false. Reasonable simplifications are thought to be true more often than not; they may ignore minor complications, but they identify the most consequential features of reality. False assumptions may be correct in some situations—a stopped watch is right twice a day—but usually they are incorrect. They may focus on conditions that matter at the margins, but they ignore the most important determinants. I fail to understand what useful insights can emerge from assumptions that are known to be false. Clarke and Primo are right about that much: passing a test says nothing about the truth of a model if its assumptions are false. However, this is not likely to happen unless the testing is flawed. Reasonable simplifications are more likely to lead to predictions that survive rigorous testing.

A good motto would be: Test Everything! Test assumptions, test predictions, test indicators for validity and reliability. As we test, however, we need to explicitly specify appropriate boundary conditions. The third obstacle to multimethod consensus is the fact that we are not in the habit of doing this. Yet we must: the political world is a heterogeneous place. There are not likely to be many one-size-fits-all propositions that hold true for all times and places. I think the kind of theory we are developing is a kind of hierarchical theory: there is one model that works in context A, a different model that works in context B; then a meta-model that explains why each model works in each context; and so on, with models nested inside meta-models nested inside meta-meta-models, which taken together constitute a theory. This is very similar to Bennett and George’s notion of “typological theory” (George and Bennett 2005).

If this is correct, then I see two tendencies that inhibit the development of this kind of hierarchical theory. One is acting as though we did not live in a political world of complex conditionality. No one really believes this: that there are any non-trivial statements about politics that are true anywhere and anytime. Nevertheless, when we test propositions by building databases that indiscriminately pool absolutely all the observations that we can find, we are acting as though there really is unit homogeneity: the same causes have the same effects in all possible contexts. To correct this tendency, we need to think very carefully about the non-arbitrary conditions under which our hypotheses should be expected to hold. We must also state these boundary conditions explicitly.

The opposite tendency is to take it for granted that everything is so complex and conditional that there is no point in trying to generalize. This leads to a narrowing of theory or even deemphasis of theory, and to myopia: the tendency to ignore large, slow-moving, structural causes and to give too much weight to small, dynamic, proximate causes. The remedy for myopia is exactly the same: thinking very carefully about the conditions under which our hypotheses should be expected to hold, and stating these boundary conditions explicitly. When we do this, the boundary conditions give us the cues we need to link the findings of case studies with the findings of large-sample analysis.

All of these problems are difficult because they involve breaking old habits, developing new skills, or exposing our cherished ideas to great risk of falsification. Nevertheless, I am optimistic about the future of multimethod research. It has real advantages and it is, I think, our future. All of these problems are in principle solvable. Being aware of them is the first step toward a better understanding of politics.

Note

1 Clarke and Primo (2007) would object to the very notion of a model being true or not. They hold that models are merely objects; they are neither true nor false. Their predictions may be true in some contexts and false in others, so the goal of political scientists is to determine what each model is useful for. In my view, models that make false predictions about phenomena they intended to explain, or that make predictions that are usually false, are themselves false.

References


Does It, Really? Measurement Error and Omitted Variables in Multi-Method Research

David Kuehn
University of Heidelberg, Germany
kuehn@uni-heidelberg.de

Ingo Rohlfing
University of Cologne, Germany
rohlfing@wiso.uni-koeln.de

A recurring tenet in much of the recent discussion on combining case studies and regression analyses and, to a lesser degree, case studies and QCA, highlights the purported ability of Multi-Method Research (MMR) to overcome the classic problems attached to each of the individual methods. The large-N part is employed to uncover cross-case regularities and causal effects, while in-depth case studies are undertaken to identify causal mechanisms in a subset of cases. In this, one purpose of case studies is to cross-validate the insights of the large-N analysis by identifying possible measurement error and exposing potentially omitted variables. Thereby, it is assumed, case studies are able to make the large-N analysis more robust.

In this essay, we critically revisit these arguments from a methodological perspective. We argue that the ability of case studies to effectively enhance the inferential quality of the large-N method is significantly limited due to the very problems that they are supposed to solve: measurement error and omitted variables. While there is some discussion of these two issues on the cross-case level (King, Keohane, and Verba 1994; Lieberson 1991), there is only limited and largely implicit reflection on their relevance for the within-case level. This is unfortunate, as measurement errors and omitted variables on the within-case level are no less damaging for the inferential power of case studies as they are for large-N analyses. Measurement error and omitted variables are particularly likely to occur in process tracing because of problems in the use of sources for data collection and the generation of inferences on the basis of this data. We, therefore, maintain that the ability of the case study to serve as a double check on the quantitative part of MMR designs is overstated. To be sure, many of the issues we highlight in this paper have been thrown up in the methods debate before. However, we think there is an insufficient transfer of the arguments made in the realm of the case study literature to the work on MMR, which is what we aim to achieve in the following.

This essay proceeds in four parts. In the first section, we shortly revisit the methodological literature to extract the two main functions of case studies as robustness checks in MMR designs: the identification of measurement error and the uncovering of omitted variables. In parts two and three, we discuss to what extent process tracing can deliver what the MMR literature promises. The fourth section concludes with a short summary of the argument.

The Role of Case Studies in MMR

Given that there is a variety of possible methodological approaches which can be integrated into a single research design (cf. Johnson, Onwuegbuzie, and Turner 2007), most of the recent discussion of MMR—our essay included—is concerned with the combination of statistical techniques, and less often QCA, and in-depth case studies (e.g. Bennett 2002; Lieberman 2005; Schneider and Rohlfing 2009). In these designs, the case study part serves two related purposes. Its primary function is to establish a causal relationship by elucidating causal mechanisms. This role addresses the well-known caveat that a cross-case regularity does not necessarily reflect causation and mirrors the increasing consensus on the relevance of mechanisms for making causal claims. Observed cross-case regularities do not provide a solid foundation for inferring causality unless it can be empirically demonstrated that the purported cause is linked to the outcome via an uninterrupted process (George and Bennett 2005; Hall 2006). Case studies are claimed to provide this foundation by tracing the processes through which the causes produce the observed effects. The second function of case studies is to cross-validate the specification of the cross-case model by checking for measurement error and omitted variables, thereby improving the robustness of the large-N results (Fearon and Laitin 2008; King, Keohane, and Verba 1994: 152–183). In this essay, we focus on the latter issues of measurement error and omitted variables on the within-case level. The issue of identifying causal mechanisms through case studies is dealt with more extensively in another paper of ours (Kuehn and Rohlfing 2009).1

Measurement error is understood as the use of indicators which are either not reliable or lack validity in capturing the empirical content of the underlying concept (Brady and Collier 2004: 295). Validity problems may derive from mistakes in concept formation, for example when important attributes and, thus, indicators are missing or redundant, or from flawed measurement, for example if an indicator with low content validity is used (Adcock and Collier 2001). The dangers of measurement error for the robustness of statistical findings are well documented. Depending on whether measurement error is systematic or non-systematic, whether it occurs on the dependent or the independent variable, and whether it is correlated across variables, the estimates can be rendered inefficient, biased, and inconsistent (Herrera and Kapur 2007). QCA suffers from severe problems as well when measurement error is present (Seawright 2005). To some extent, the problem of measurement error is inherent to social science research simply because the concepts used to systematize the empirical world are constructs made by researchers. Accordingly, there will always be some degree of imprecision involved in fitting theoretical constructs to the empirical phenomena of interest (King, Keohane, and Verba 1994: 152).

In large-N research, there are statistical means by which measurement error can be tested and corrected for (Rabinovich 2000). The MMR literature, however, suggests in-depth analysis of one or a few number of cases as an additional check for
the quality of concepts and the validity of indicators (Coppendge 1999; McKeown 1999). It is to note that the reliability of an indicator cannot be evaluated through case studies as this would require repeated measurement of the same concept in more than one or a few cases. As regards concept formation, it is claimed that case studies can provide valuable information on construct validity. This can be done, for instance, by paying attention to context and thus avoiding conceptual stretching (Adcock and Collier 2001). Concerning the scoring of indicators, process tracing is held as less prone to measurement error, “because it can intensively assess a few variables along several qualitative dimensions, rather than having to quantify variables across many cases” (George and Bennett 2005: 220). In sum, the ability of case studies to provide superior concepts and measurement has been identified as one of the most important contributions that method has to offer (King, Keohane, and Verba 1994: 151–168).

Similar arguments are made for the problem of omitted (or missing) variables, which undermines the cross-case analysis in QCA and produces biased estimates in regression analysis (Seawright 2005). As is the case for measurement error, there are statistical tools to check for the existence of omitted variables (Clarke 2005), but case studies are argued to be a particularly suitable tool to uncover hitherto neglected factors in the cross-case analysis (George and Bennett 2005: 34). Furthermore, as the range of possible confounders is in principle unlimited (King, Keohane, and Verba 1994: 174), case studies are supposed to identify exactly which variables should be included in the regression model. In this, both typical and deviant cases are deemed to be particularly useful for the search of omitted variables. Process tracing in typical cases is undertaken to check for spurious cross-case patterns, that is, the working of a third variable which influences both the independent and dependent variable. The analysis of deviant cases can help to identify variables which can then be included into the model in order to enhance model fit (Lieberman 2005).

Measurement Error in Within-Case Analysis

Thus far, the literature discussed the implications of measurement error with respect to cross-case inferences only (King, Keohane, and Verba 1994; Lieberson 1991). Measurement errors are usually attributed to the researcher’s “slip of the hand,” that is, random mistakes in coding the cases, or to the use of weak concepts or invalid indicators. Besides these general arguments there are, to our knowledge, no guidelines of how actually to perform a within-case analysis in order to improve concepts and measurement. We argue that the rather optimistic view on the potential of process tracing is not warranted for two reasons. First, the case study suffers from a variant of the classic small-N problem because the quality of a concept and measurement must be assessed in a larger set of cases. Second, within-case analysis is similarly prone to measurement error as large-N analysis, which undermines its utility for cross-validation.

Concerning the first problem, process tracing may in fact show that a concept is misspecified or that a weak indicator has been used. Intimate knowledge of cases provides the ground for potentially more realistic interpretations and more valid classifications of empirical facts, for example by identifying functional equivalents which would go undetected in the standardized coding procedures typical for large-N studies (Adcock and Collier 2001). A comparative case study interested in political opposition, for instance, may suggest that in some countries this concept is better measured through the behavior of parties and that in other countries interest group politics might be a more appropriate indicator (Verba 1967). Ultimately, however, the validity of the indicators must be evaluated by calculating their convergent or discriminant validity or, more preferably, construct validity (Adcock and Collier 2001). These tests are indispensable because concepts or indicators are not supposed to fit one or a small set of cases particularly well. Instead, conceptualization and measurement should be as widely applicable as possible to the population of interest (Carmines and Zeller 1979). If, for example, a given concept does not fit some individual cases but is valid for a large share of a given population, it would be of little use to reconceptualize it in order to make it fit the deviations. If at all, it would be considered more plausible to eliminate these cases on theoretical and conceptual grounds. At the same time, however, the latter strategy runs the danger to ad hoc delineation of the population. There probably are always cases for which the concept or indicators are not optimal, but the lack of fit does not warrant their exclusion.

Process tracing employed for improving concepts and measurement, therefore, is confronted with the same small-N problem known from causal inference (cf. Lieberson 1991): even if one finds that the concept does not fit well one or a few numbers of cases it is not clear which conclusions should be drawn without knowledge about the generalizability of what one learned from the cases under scrutiny. Of course, qualitative within-case analysis is particularly valuable both for causal inferences and the improvement of concepts and measures with respect to the cases under analysis, that is, internal validity. However, the degree to which these insights are externally valid is impossible to gauge with much certainty because such an evaluation would include statements about cases which have not been subject to empirical analysis.

The second neglected problem of case studies is measurement error on the within-case level. Measurement means assigning scores to indicators, which is not necessarily as unproblematic in small-N research as some MMR researchers seem to believe. Measurement error in case studies can derive from two sources: the deficient collection of empirical data and the incorrect interpretation of this data. Concerning the first problem, it is well known that inferences drawn from process tracing are necessarily only as good as the sources on which they are based. The specific problems typical for different sources of empirical evidence, e.g. primary and secondary sources, newspapers, and interviews, have been extensively discussed elsewhere and need not be elaborated upon here (e.g., Lustick 1996; Thies 2002). The usual recommendation to cope with these problems is triangulation, that is, to draw on multiple sources, thereby addressing their respective shortcomings. Case studies are supposed to be particularly suited
triangulate data as they offer "an opportunity to fact check, to consult multiple sources, to go back to primary materials, and to overcome whatever biases may affect the secondary literature" (Gerring 2007: 59). Triangulation certainly may help to ameliorate the problems of data collection in process tracing, but a certain, and often considerable degree of uncertainty about the full picture remains. It is exactly this uncertainty from which measurement error in small-N research derives. For illustrative purposes, assume that one is interested in the prevalence and severity of human rights violations in order to code the democratic quality of a country. Due to the very nature of many autocratic regimes, it will be difficult to obtain sufficient data to effectively measure the degree of human rights violations in those countries as official documents will be hard to obtain, newspapers and media reports will likely be subject to censorship, and interviewees will often be reluctant to share information with foreign researchers for fear of repression and retribution. In fact, the problems related to validly measuring democratic quality are systematically related to the concept itself. In this example, it is likely that the available empirical evidence biases the classification of the regime, possibly suggesting a better democratic quality than the actual state of civil rights would actually allow. The difference between the observed and the true state of democratic quality in that country, however, cannot be validated based on the available empirical evidence and thus it is not possible to eliminate the existing measurement error. In other examples, the limited availability and inherent bias of sources might not be as evident as in this example and measurement error might occur without the researcher being aware of it.

But even if sufficient empirical evidence is accessible to the researcher, measurement error on the within-case level might just as well derive from subsuming a certain observation under a wrong concept or interpreting some piece of empirical evidence wrongly. This danger might be at least as prevalent in process tracing as it is for cross-case research. First, empirical observations on the within-case level are usually highly disparate in nature and are not typically easily compared with each other. Second, and in contrast to observations on the cross-case level, the rules of how to identify, interpret and code within-case observations are usually not formalized, standardized and made explicit (Collier, Brady, and Seawright 2004; King and Powell 2008). Therefore, the replication of measurement in order to uncover measurement errors in within-case analyses is often hardly possible.

As an example, Grieco (1990) performs a deep and rich case study of how some agreements of the GATT Tokyo Round were implemented in the 1980s. He derives multiple observable implications from two competing theories, neo-realism and neoliberalism, and claims that the empirical evidence is more in accord with the neo-realist explanation. However, while the empirical within-case evidence he provides supports the neo-realist argument, it is not conclusive as the process observations he cites can equally well be subsumed under the plausible competing hypothesis that trade negotiations are actually shaped by domestic politics and the interests of economic actors. For instance, Grieco quotes a report of an EU commis-

sion saying that "the 'cost of non-Europe' represents a direct, serious threat to essential components of the Community's industrial potential." Without knowing whether the "Community's industrial potential" is deemed important for the success of Europe, which is what neo-realism predicts, or whether this statement simply reflects the influence of strong lobby groups aiming to operate in a large, liberalized market, this piece of evidence is inconclusive and subsuming it under one of the competing accounts runs the risk of committing a measurement error. The problem of misinterpretation of process evidence is particularly salient for those within-case hypotheses which "bottom out" (Machamer, Darden, and Craver 2000) on the level of individual or collective actors and try to explain social phenomena by these actors' intentions. Njølstad (1990), for instance, shows that such problems are ubiquitous even in a well-researched area like US nuclear policy which is replete with disagreements on how to interpret the supposedly well-documented interests and beliefs of even the most important actors and how these may have contributed to even the most critical and obvious events and developments.

Omitted Variables in Within-Case Analysis

Similar rationales hold for the problem of omitting variables. It might be that the researcher is unable to identify a relevant factor due to sheer lack of information in the available sources. Since the researcher usually is not able to decide if she has collected a sufficient amount of evidence, the existence of one or more highly important pieces of evidence may go undetected (Thies 2002). Based on limited within-case evidence, a researcher could fall for a type I error by inferring a causal process between X and Y, while in fact the evidence is flawed or there is some hidden evidence that the actors concealed their true motivations. At the same time, there is a risk of a type II error by erroneously accepting the null hypothesis when indeed there was some causal process which remained undetected because of unavailable evidence.

Either way, the absence or presence of seemingly convincing within-case evidence does not necessarily mean that the identified causal process (or its absence) is true. Assume, for instance, that one wants to test two competing explanations of the democratic peace thesis on the within-case level. If the empirical evidence is compelling for one account and weak for the other, one would infer that process tracing corroborates the first explanation and disconfirms the second one. However, if these inferences and scores are primarily grounded in the accessibility of evidence in favor of the first explanation, one would draw erroneous conclusions. This could be, for example, the case when one explanation focuses on civil society activities in democracies and the other hypothesizes peaceful effects of economic ties between democratic countries. Protests of civil society would be quite visible because their actions take place in public. At the same time, however, there may be massive lobbying behind the scenes by economic interest groups that fear to lose from war. Furthermore, from the perspective of politicians, it might be more attractive to launch a peace initiative by referring to civil society and democratic norms instead of economic interests and pressure.
group lobbying. In this example, process tracing is more likely to uncover empirical evidence in favor of the civil society explanation and an important variable would be omitted from the theoretical account. However, if politicians and economic actors in question conceal their true motivations in interviews, and if primary sources are kept secret, one cannot do more than speculate about the actual role of economic interests.

Conclusion

In much of the recent MMR literature, case studies are promoted as adequate tools to check the regression results for measurement error and omitted variables. We have argued that the method might not be as capable to fulfill this function as it suffers from similar weaknesses as large-N methods when it comes to handling inferential challenges. As a particularly pressuring problem we have identified the quality and use of sources. They affect the possible contribution case studies can make to overcome measurement errors and to identify omitted variables. Therefore, the opportunity of within-case analyses to cross-validate the large-N part is limited. This does not mean that qualitative within-case analyses cannot improve one's trust in the large-N findings at all and that it therefore should not be undertaken for these purposes. However, we want to caution against putting too much hopes in the insights gathered from process tracing and recommend to additionally rely on the established diagnostic tests for measurement, omitted variables, and other specification issues in regression analysis (Rohlffing 2008).

Notes

1 By focusing on the second function of case studies we are, however, able to make inferences about the first function because they are closely related. If we can show that the case study part in MMR designs is not able to test the robustness of the large-N analysis, the method will neither be able to fulfill the even more demanding task of founding causal processes.

2 No such means are available for QCA at present.

3 Provided that the omitted variable is correlated with the included variables, which is almost always the case, the consequences of omission are ambiguous. It is only in the bivariate case that the estimate is biased toward zero (King, Keohane, and Verba 1994: 171).

References


Fuzzy logic at its origin (Zadeh 1965; see Kosko [1993]) or McNeill and Freiberger [1994] for very accessible introductions) was created in part as a mathematical theory of semantics. As such it is an appropriate tool for transforming raw data and indicators into numbers that better match the theory and meaning of key theoretical concepts.

In this essay I introduce the Fundamental Principle of Variable Transformation. This principle requires that all transformations of variables be meaning preserving or increasing. To use one of my examples below, the principle requires that if one logs GDP/capita it should better represent what the scholar means by, say, wealth than the untransformed data. From the qualitative perspective all transformations which do not conform to this principle are suspect.

**Meaning Transformations versus Statistical Transformations**

Befitting the two-cultures metaphor, the language used to talk about the “same problem” varies significantly across the traditions. Within the quantitative culture one speaks of “variables” (aka concepts) and “indicators” (aka numeric data). Typically, the nature of the relationship between variables and indicators is not discussed. For example, is the relationship between the indicator and variable causal? Is it merely correlational? In practice, one often laments that the indicator does not reflect, measure, or represent well the theoretical variable.

Within the qualitative culture one would ask about the semantic, theoretical relationship between the numeric data, say GDP/capita, and a concept such as economic development. What is the definition of economic development, and which attributes of this concept are measured by GDP/capita? How do different levels of GDP/capita relate to the concept of economic development?

Thus, both qualitative and quantitative cultures can take the same data, e.g., GDP/capita, as somehow a measure or related to wealth, but they then proceed to do quite different things. Within the statistical culture GDP/capita as a measure of wealth is commonly transformed, e.g., logged, for statistical reasons. A qualitative scholar will take the same data and ask about the theoretical and semantic relationship between GDP/capita and the concept of wealth. Fuzzy logic is then a means to transform, or calibrate in Ragin’s terminology, these raw data to better reflect what the theorist means by wealth.

**Standardization and Concept Meaning**

Perhaps one of the most popular transformations is to standardize a variable. In some fields it is quite common, if not virtually obligatory, to standardize variables before performing statistical analysis. Standardization often does not change the statistical results, because most parameter estimates retain their statistical properties such as unbiasedness when the variable is subject to a linear transformation. To choose an example, perhaps familiar to members of the QMQR section of APSA, I take Gerring’s (2005, chapter 5) advice to standardize variables in order to select case studies based on their “extreme” values. Gerring uses the concept of democracy as coded and measured by the polity project (Jaggers and Gurr 1995).