Symposium: Field Experiments and Qualitative Methods

Battling Onward: The Debate Over Field Research in Developmental Economics and its Implications for Comparative Politics

Edmund J. Malesky
University of California, San Diego
emalesky@ucsd.edu

Victory has finally been declared after a fierce internecine struggle, but the conflict will never be recorded in the Correlates of War dataset. In fact, except for a few notable exceptions, news of the battle has hardly seeped into political science journals at all; and its impact has certainly not been appreciated by the group best positioned to capitalize on the terms of victory—comparative scholars with deep country knowledge.

The war has been fought on the pages of economics journals[1] and the main battle lines have been drawn over the standards of admissible evidence used to assess theories of economic development. On one side were practitioners of old-school research, who predominantly studied economic growth, trade, and capital flows, and for whom the major research tool has been large-*N*, cross-national regressions (Rodrik 2008). On the other side were practitioners of the New School of Developmental Economics (NSDE). These researchers, sometimes referred to sardonically as "randomistas," abhor cross-national designs, distrust high-tech statistical band-aids, and believe the only evidence that is worth paying attention to are randomized field trials, where the impact of treatments can be readily observed, or natural experiments, where accidents of history mimic randomized trials (Banarjee 2007). Although there is still some dissent among those whose research projects do not easily lend themselves to these methods, NSDE tools are now considered to be the cutting edge of economics research; their work dominates major economics publications and they have been major beneficiaries of international and federal research funding.[2]

The logic of the NSDE research agenda can be applied to a number of research arenas that political scientists are interested in as well. As I will argue below, comparative politics is a particularly target-rich environment. Moreover, NSDE approaches significantly alter the entry costs of empirical work. Mathematical facility is relatively less important, while a premium is applied to detailed country knowledge that allows researchers to identify appropriate natural experiments and use contacts to achieve buy-in on conducting gold-standard policy interventions. In short, comparative political scientists are better-suited than other sub-fields to begin applying these techniques in our work. In this essay, I introduce the NSDE, discuss their most prominent methods, and assess pros and cons of applying this approach to fieldwork in comparative politics.

Tools of the New Developmental Economists

NSDE congratulates their forefathers in developmental economics for developing a large number of important theories, but they argue that these theories have been tested very poorly, if at all. First, they argue that large-*N*, cross-national studies are insufficient for addressing the detailed micro-logic implied by economists' formal models. Why focus our research on aggregate associations, when a theory specifically involves individual behavior? For instance, an argument that links property rights development and economic growth implies that individual businessmen are expanding investment because they feel more secure. Thus, an exogenous shock to property rights allocation should be followed by greater investment at the firm-level. It is this firm-level behavior that should be the focus of investigation.

Second, NSDE practitioners argue that measurement of key theoretical concepts has traditionally been slapdash and that better fieldwork can offer more precise measurement testing. Ray Fisman's 2001 study of the impact of political connections and business performance in Suharto's Indonesia was a landmark example of the dividends of concern for careful operationalization. Fisman demonstrated the importance of political connections by studying the performance of listed companies on days that the media reported news about Suharto's failing health. He found that companies closely associated with Suharto, his family members, or high-ranking Golkar party members dropped precipitously after such events, when compared to other listings.

Third, they argue that their predecessors have been sloppy about *identification*, by which they mean resolving endogeneity and selection bias in their empirical tests. It is worth dwelling on what Abjhit Banerjee (2005: 31) has termed the "sacrifices to the harsh god of identification," because it is really the motivating principle of NSDE, the flag around which they all rally. Their ferocity in identification stands in sharp contrast to the cavalier approach with which it has been regarded in comparative politics to date.[3] King, Keohane, and Verba (1994), of course, described the problem of endogeneity in their seminal book on methodology in social science, but their main take-home rule was to avoid it in designing a research project. This can be less than helpful to new graduate students, because many of the big theories of comparative politics that fascinate us have some element of endogeneity. We are therefore forced to: (1) ignore the critical questions of the discipline, which rarely makes for a successful dissertation; (2) wave our hands, declare "everything is endogenous," and push onward with research project that is flawed at its core; (3) lag our key causal variable in regression analysis and claim that this accounts for reverse causality, while praying that reviewers do not call us on the sleight of hand; or (4) use a mixed-methods approach that uses case studies to process

trace the true sequencing of causes and outcomes (Gerring 2007: 37). None of the first three options would past muster with NSDE and option four can become easily bogged down in debates about whether the case selection was appropriate for the question, generalizable beyond the specific instance, and avoids biased historiography in the choice of sources.

For NSDE, adequate identification strategies require one of three techniques, which I list in order of their favor by the field:

Field experiments, where the researcher treats a randomly selected sub-sample with a particular policy and therefore knows for certain that policy was exogenous.[4] An excellent example of this type of work is Ben Olken's 2007 analysis of corruption in Indonesia. A long-held theory in the discipline and among practitioners was that decentralization of responsibilities to subnational governments, combined with participatory oversight mechanisms (e.g., town-hall meetings) limited corruption.

Previous economists and political scientists studied this question, but the primary analysis was to regress a cross-national indicator of corruption on cross-national indicators of levels of administration decentralization (See Rose-Ackerman 2004 for a review of this work). Olken was not satisfied with these results because the cross-national measures of corruption perception (often by foreign investors) were not precise enough to link them directly to the decision to decentralize. Alternatively, organizations like the World Bank (2004) had developed case studies of grassroots accountability mechanisms, but these studies were plagued by selection bias. Sites had been specifically selected due their good governance relative to their peers.

Unsatisfied with previous work, Olken put this theory to the test by convincing the Indonesian office of the World Bank to randomize its oversight over six hundred village-level road-building projects. All villages received a similar pot of money for road construction, but Olken randomized three accountability mechanisms. Some villages were trained in the art of town-hall meetings and local elections to the town council. Other villages were threatened with an audit by a central regulatory agency. And a third group was not provided with any accountability mechanism at all.

Olken then built test roads in order to determine how much the materials cost to build a kilometer of road. Using a team of engineers, Olken then hired a team of engineers to take core samples of each of the 600 village-level road projects. He found that local participation through town councils was not significantly different from no monitoring at all. Threat of a central audit, however, did significantly reduce theft of road materials.

Olken's work undermined a key argument in theories of corruption monitoring. It is a critical finding that continues to reverberate in development circles. What made it possible was Olken's detailed understanding of Indonesia that allowed him to identify the most precise way to measure corruption and to identify organizations that were willing to have their policies evaluated.

Discovery of an ideal and verifiable natural experiment that provides an exogenous shock to the key causal variable. Sometimes, despite a researcher's best persuasive efforts, it is impossible to convince governments and donors to randomize a policy treatment, especially in the areas that concern political scientists. In other cases, career motivations prevent government officials and donors from accepting randomized evaluations of their interventions. In these cases, I have found the new developmental economists to be at their creative best in identifying shocks which function as natural experiments, creating an exogenous and quasi-random treatment that can be exploited (Dunning 2008b).

My favorite example of this is the brilliant work on child soldiering by Blattman and Annan (2006), because they tackle an incredibly complex issue that is of direct interest to comparative political scientists: what are the economic and psychological consequences of child soldiering? Selection bias poses a clear problem for this work; perhaps the same traits that make a child attractive for abduction into child soldiering also impact their chances of post-war success. Or perhaps militias focused their abductions on particular ethnic groups, which were already disadvantaged economically.

To resolve this problem, Blattman and Annan explore the methods of abduction, discovering that they primarily occur through raids on isolated minority villages. As a result, the choice of children is haphazard, depending solely on who could not manage to escape at the time of the raid. To construct their sample, they select abducted and unabducted siblings from the same household, which allows him to hold constant most demographic features that would confound analysis. Using this creative strategy, the authors find that child soldiers are less likely to finish elementary education, achieve functional literacy, and obtain skill-intensive labor in later life.

Occasionally, scholars dig deep into history to identify an appropriate natural experiment. Banarjee and Iyer (2006), for instance, use British patterns of land administration in colonial India as an exogenous treatment for institutional development. Different British colonial administers implemented either a landlord based system or a land tenure system in different regions of India. They find that that original treatment leads to vast differences in electoral participation, public goods provision, and ultimately economic growth.

Once again, these findings do not involve mathematically complex approaches; it is the area-studies knowledge of the researchers and their passion for particular issues that allowed them to exploit the opportunities that were presented them.

Discovery of the ideal instrumental variable, which can be employed in a two-stage procedure to resolve endogeneity bias. The worst-case scenario for NSDE is to fall back on an instrumental variable. Here, randomization to address endogeneity and unobserved heterogeneity has essentially been taken out of the scholars hands and a suitable exogenous treatment (that simulates random assignment) cannot be identified. Usually, a scholar resorts to this technique when the question is considered to be of critical importance to economic theory or policy choice. In these cases, the goal of instrumental variable selection is essentially to simulate a natural experiment by identifying the exogenous portion of a key causal variable and testing the impact of that exogenous com-

ponent on the outcome of interest.

Instrumental variables are common in political science as well, but what differentiates the developmental economists is their explicit consideration of IV-regression as a substitute for a natural experiment and the care they give to interpreting the results (well…sometimes).

By far, the most famous exemplar of this type of analysis is Angrist's (1990) study of the impact of military service on lifetime earnings. Of course, the choice of a military career is not exogenous, so disentangling the factors that lead someone to choose military over civilian employment is nearly impossible using the data available on social security records. Angrist, however, knew that a great deal of service in Vietnam took place as a result of a draft-era lottery system where soldiers were selected by birth date. Because a lottery is by definition random, Angrist could use birth date as an instrument to identify the effect of military service, finding significant and negative results.

Most applications of instrumental variables would have stopped there, but Angrist created a new benchmark for the field by defining the Local Average Treatment Effect (LATE) of his instrumental variable selection. Because Angrist and the New Developmental economics school see instrumentation as a poor man's substitute for randomized treatment, they explicitly seek to identify the parameters of their imperfect treatment. Angrist recognized that the birth-date lottery was a treatment that did not apply to all the observations in his sample. Specifically, two groups were excluded: soldiers who self-enlisted in military service and draft dodgers who shirked their randomly selected responsibility. Thus, he stated unequivocally that his findings cannot be applied to these sub-samples; they are limited to the general populace who may or may not have received the lottery treatment.

The concept of LATE is often extremely helpful for understanding why point predictions and the impact of control variables can change dramatically in an IV model, even if the instrument selected is strong and meets the exclusion criteria. The LATE allows readers to understand the sub-population to which the instrumented treatment applies.

Another excellent example of the careful NSDE approach to instruments is Woodruff and Zenteno (2007). In this clever piece, the authors use distance from the railroad in Mexico to identify migration and consequently study its impact on micro-enterprises in Mexico. The railroad instrument works because Woodruff and Zenteno are able to show how proximity to a railroad station incentivized the decision to migrate, but had little impact on the choice to start a new business. Once again, the two scholars nail down the LATE, so the reader understands the parameters of the analysis.

For all three above techniques, panel models that study the same observations over time are requisite; single-shot surveys and data collection are uniformly thought of as inappropriate for identification. Only through panels can a researcher truly identify the hypothesized changes in time resulting from the treatment. At the very least, scholars should try to construct retrospective panels from existing surveys, but these are imperfect to the memory deficiencies of respondents.[5]

Can these Techniques be Applied Successfully in Comparative Politics?

Absolutely; in fact, it is already beginning to take place in a number of interesting arenas. A recent issue of the *American Political Science Review* (102:1, 2008) featured three field experiments and a survey experiment among its ten articles. In comparative politics, Donald, Green, James Gibson, Daniel Posner, Jeremy Weinstein, Marcartan Humphries, Leonard Wantchekon, Susan Hyde, and Elizabeth Paluck have been at the forefront of using these NSDE approaches to study explicitly political questions.[6]

But there are still many more opportunities. The same issues which spurred on the new developmental economists are currently present in political science. Just like the economics literature, comparative politics has a range of important theories that are built upon the behavior of individual actors, but have only been tested in aggregate cross-national analyses. As with any new approach, there are pros and cons to adoption. The individual actions of voters, interest groups, and political entrepreneurs buttress a number of the cross-national associations that have been discovered in our discipline. Whether these actors really behave in the manner assumed remains an open question.

The Benefits of a Developmental Economics Approach for Mixed-Method Researchers

The most important benefit of following developmental economists in their approach is the precision of the analysis. Scholars are removed from the well-known difficulties of studying large cross-national datasets, where measurement error, unobserved heterogeneity, and endogeneity frustrate even the most theoretically sound projects. Political science has recently been at the forefront of devising new and more sophisticated approaches to dealing with unsatisfactory data. Our graduate students today must effectively minor in statistics in order to address the myriad problems hidden in datasets such as Polity IV or the World Values Survey. But as Donald Green (2005) has pointed out, observational analyses often perform quite poorly relative to experimental analyses of the same issues, even when observational scholars use cutting-edge statistical specifications.

For NSDE, the costs of analysis are on the front end. If a field experiment is well designed or a natural experiment carefully chosen, a scholar can analyze the resulting data with comparably easy statistical operations. In fact, simpler is generally considered to be better. A sure-fire way to engender suspicion in an NSDE audience is to present an empirical result that shows up after a sophisticated procedure (i.e., error correction models or generalized method of moments (GMM) regression) that does not hold using simpler specifications. Simplicity also has the dual advantage of allowing the research to be accessible to non-quantitative scholars, which should go a long way toward uniting our divided field. Particularly because, as Elizabeth Paluck argues in her contribution to this symposium, experiments generate excellent qualitative as well as quantitative information.

A second benefit is that it plays to the strengths of comparative political scientists. We often know these countries better than anyone else, speak the local language, and possess a range of contacts in government, business, and development sectors. Often, our fascination with particular countries is what drove us to study comparative political science in the first place. Because of this, we are ideally suited to identify appropriate national experiments or convince practitioners (policy makers or donors) to randomize a policy intervention. It is ironic that we have surrendered this ground to the "number-crunching" economists.

We shouldn't be dismayed that economists got the jump on us, however, because many of their randomizations allow for what Green has termed "downstream experiments," where the same randomization can allow for follow-on studies (2005). Green points out that educational attainment may have been initially studied by economists—through randomly distributed college scholarships—to study its impact on variables of economic interest, but we can return to this sample later on to study education's effect on election turnout, racial tolerance, civil engagement, or job performance.

The Negative Side of NSDE

Before drinking the NSDE Kool-Aid, it is worthwhile to highlight some limitations of the approach. NSDE scholars actively discuss these issues, of course, but there is still work to do in resolving them.

First, a focus on the techniques may force scholars to avoid big questions. As discussed above, the ideal target for NSDE tools in comparative politics is when a large theory implies specific individual-level behavior. Creative scholars can focus on the micro-logic and identify a test of it. In KKV (1994) terminology, they can test the individual observable implications of the larger theory. If the theory is an explicitly macro theory about the behavior of states in the world system or parties within political systems, cute identification strategies may not be available. Rare is the government that allows a researcher to randomly sow ethnic strife across subnational entities. And domestic terrorist organizations, a hot topic of current scholarship, are too few in number and too strategic in their choice of targets to see if a natural experiment is applicable.

NSDE scholars would certainly retort that my characterization is unfair and that many macro-theories have individual implications if you think carefully about the design. For instance, instead of focusing on states as the unit of analysis, one could focus on specific policies, legislative documents, or exploit subnational variation. Yet even diehards would concede that there are certain issues that do not lend themselves well to gold-standard randomization. Here, the best a scholar can do is hope to identify an instrumental variable for the task at hand. And, well…this can sometimes be like waiting for Prince Charming to arrive. The instrument must meet some very high standards (known as the exclusion criteria), must be suitably strong, and must survive a range of challenging (and exhausting) diagnostic tests.[7] And that is just for cross-sectional analysis. In a panel model, Prince Charming must vary over time! What is the conscientious scholar to do?

A second problem is the generalizability of the research findings. NSDE scholars unapologetically "go small" in their research designs. Rather than controlling *ex post* for confounding factors, they select them *ex ante* to minimize observed heterogeneity, hoping their randomization procedures will address unobserved heterogeneity. But how much analytical leverage do we really get from a few hundred villages in Indonesia or India? This criticism is very similar to those leveled against detailed case-study work. How can we be sure that the results of a theoretical test will travel to other developing countries, or even developed countries (where field experiments are rarely attempted)? Quite simply, we don't; and the vagaries of academia mean we probably never will. Clever field experiments and natural experiments get honored only once in a top-shelf publications. Applications of the design to other contexts (normal science in a Kuhnian sense) are simply not rewarded by the field.

NSDE chooses to privilege small studies executed to near perfection with limited external validity over large, observational studies that can never be executed perfectly but have some external validity. As Banarjee (2005) puts it, "…even if we have many low quality regressions that say the same thing, there is no sense in which the high quality evidence becomes irrelevant—after all, the same source of bias could be affecting all the low quality results." In essence, Banarjee believes the inference problems from poorly designed, large-*N* works are so severe that the external validity they purport to offer should be discounted. Actual learning, however incremental, is preferable to the illusion of learning that observational studies offer (Green 2005).

Third, banking on an NSDE technique early in an academic career can be dangerous, because the choice to use them is not always in scholars' hands. In the case of field experiments, government officials and donors must agree to a randomized policy intervention, which does not always coincide with their own career incentives. It can be much easier to pre-select areas where one knows a controversial policy will work than randomize and remove all doubt. There is no doubt that fewer of my e-mails to Vietnamese and Cambodian practitioners are returned since I began pushing randomized approaches.

Similarly, natural experiments may not occur exactly where a research question demands they do. This can create an odd form of inductive research, where a researcher identifies an exogenous shock and then works backwards to identify a question that can be answered using it. I think we all can agree that is not a practice that should be condoned by our discipline. Scholarship advances with the work of creative puzzle solvers, not puzzle hunters.

Finally, I have one word of warning for scholars who are fortunate enough be able to use an NSDE technique for their research question. Randomization and exogenous shocks do not excuse sloppiness on other aspects of the research program. The vast majority of experiments rely on some form of survey to measure the dependent variable. In studying some of these survey instruments in preparation for a course on

research methods, I was shocked to discover that many of the same issues which contaminate general survey research can be found in these well-designed studies. Double-barreled questions, ambiguous terminology, and potential framing effects in question ordering are rife in many instruments. Biases such as these, which are often exacerbated by unclear manuals for hired interviewers, can be just as detrimental to inference as the problems NSDE techniques were designed to solve. It is amazing that scholars can spend so much time carefully identifying the perfect randomized intervention but create a messy survey to analyze it.

Conclusion

Despite my warnings, I generally believe NSDE techniques are good for the field. It is time that comparative politics also rallied around the flag of identification. There are a number of juicy targets available to explore with more careful research designs. If we do not study them, the economists certainly will. Economics long ago abandoned the idea that their discipline was devoted to the analysis of money. It is now a discipline of tools in search of questions, and comparative politics has some of the most fascinating queries.

I also feel strongly that all four problems I indentified with NSDE can be solved. More macro-theories will be felled once identification becomes mainstream in political science and more creative minds go to work looking for angles to cut into those theories. Generalizability has already been identified as a concern by NSDE, and organizations such as Dean Karlan's Innovations for Poverty and the Jameel Poverty Action Lab are hard at work chronicling randomized experiments and, recognizing that there is little academic benefit to replication, pushing international donors to replicate randomized experiments in new contexts. If big donors can be convinced, there will be many more opportunities for such research. There is already some progress. The Millenium Challenge Corporation and the World Bank Group's Private Sector Development Facilities have NSDE approaches as part of their organizational mandates. Finally, sloppy survey work can and should be rectified easily.

In short, I am positive about bringing NSDE to field research in comparative politics. Now, we just need to take the battle forward and assign more of these articles in our methods classes.

Notes

[1] And I imagine, although I have no definitive proof, in the review letters on submissions to those journals.

[2] For the official declaration of victory and the discussion of further territories to be conquered see "New Directions in Developmental Economics: Theory or Empirics? A Symposium." *Economic and Political Weekly*. August 2005. http://www.arts.cornell.edu/poverty/kanbur/NewDirectionsDevEcon.pdf.

[3] In less polite conversations, the NSDE has been referred to as the "Identification Taliban."

[4] See Duflo, Glennerster, and Kremer (2006) for a handy primer on how to implement such projects.

[5] NSDE scholars are not the only group concerned about inference. Other disciplines (education and psychology) have developed their own, less field-intensive modes of resolving these problems. These include direct matching, propensity score matching, and most recently, synthetic control methods for comparative case studies (see Abadie and Gardezeabal 2003).

[6] Why African specialists are at the forefront of this approach is a research question in its own right.

[7] For a nice discussion of these and other principles, see Dunning (2008a) and Murray (2006).

References

Abadie, Alberto and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93:1, 112–32.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80:3, 313–36.

Banerjee, Abhijit and Lakshmi Iyer. 2005. "History, Institutions and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review* 95:4, 1190–1213.

Banerjee, Abhijit. 2005. "New Development Economics" and the Challenge to Theory." *Economic and Political Weekly* (August): 30–39.

Blattman, Christopher and Jeannie Annan. 2007. "The Consequences of Child Soldiering." HiCN Working Paper #22. http://www.hicn.org/papers/wp22.pdf.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2006. "Using Randomization in Development Economics Research: A Toolkit." CEPR Discussion Papers 6059, December.

Dunning, Thad. 2008a. "Model Specification in Instrumental Variables Regression." *Political Analysis* 16:3, 290-302.

Dunning, Thad. 2008b. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61:2, 282–93.

Fisman, Raymond. 2001. "Estimating the Value of Political Connections." *American Economic Review* 91:4, 1095–1102.

Gerring, John. 2007. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Gibson, James. 2008. "Group Identities and Theories of Justice: An Experimental Investigation into the Justice and Injustice of Land Squatting in South Africa." *Journal of Politics* 70:3, 200–16.

Green, Donald. 2005. "The Illusion of Learning." *Deadalus* (Summer): 97–99.

Habyarimana, James Macartan Humphreys, and Jeremy Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?"*American Political Science Review* 101:4, 709–25.

Hyde, Susan. 2005. "Introducing Randomization to International Election Observation: The 2004 Presidential Elections in Indonesia." Dissertation, University of California, San Diego, Chapter 6.

King, Gary, Robert Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University.

Rodrik, Dani. 2008. "We Shall Experiment, but How Shall We Learn?" Prepared for the Brookings Development Conference, May 29–30.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55:3, 399–422.

Woodruff, Christopher and Rene Zenteno. 2007. "Migration Networks and Microenterprises in Mexico." *Journal of Development Economics* 82:2, 509–28.

World Bank. 2004. *Making Services Work for Poor People*. World Development Report 2004. Washington, DC: World Bank.

The Promising Integration of Qualitative Methods and Field Experiments

Elizabeth Levy Paluck
Weatherhead Center for International Affairs,
Harvard University
epaluck@wcfia.harvard.edu

Over the past few decades, a productive exchange in political science has explored the idea that qualitative research should be guided by the logic of mainstream quantitative and experimental methods (e.g., Brady and Collier 2004; Gerring and McDermott 2007; King, Keohane, and Verba 1994). Most of these discussions focus on the logic of regression for drawing inferences from observational data, setting aside experimentation as an ideal but rare path to causal inference. A perhaps unintended message of this discussion seems to be that experimentation is a method unrealistic for most qualitative research projects, and consequently, that experimentation is more naturally a quantitative enterprise. In short, qualitative researchers can aspire to use experimental logic for constructing counterfactuals and drawing causal inferences, but cannot use actual experiments.

This essay contends that experimentation, specifically field experimentation, can and should be more central to qualitative research approaches. This argument rests on claims about what field experimentation is as well as what it is not. Field experimentation *is* one of the strongest methods for inferring causal relationships in real world setting. Field experimentation is *not* inherently quantitative.

By randomly assigning units (individuals, communities, organizations) into two groups, field experiments can infer that differences between the groups are due to an intervening "treatment" (a media program, a land redistribution policy, elite negotiation meetings) applied to one group and not to the other. The key advantage of field experiments is that they draw causal inferences without invoking untestable assumptions that plague observational research about the groups' *ex ante* comparability.

The most straightforward reason why field experiments are perceived as quantitative enterprises may be found in the psychological concept of the availability heuristic (Kahneman, Slovic, and Tversky 1982). Put simply, there are few available exemplars of field experiments that incorporate qualitative methods or test questions traditionally associated with qualitative investigations, so experiments are thought of as quantitative in nature. My argument in this paper is that the lack of qualitatively oriented field experiments stems more from our inability to think outside of this heuristic than from unassailable methodological and epistemological divides.

Consider the potential of qualitatively oriented field experimentation using a recent outstanding set of field experiments on gender and political leadership. Chattopadhyay, Duflo, and colleagues (2004; Beaman et al. 2008) capitalized on a policy experiment in India in which the government randomly reserved one third of village council head positions for women (i.e., in councils randomly chosen for a reservation, only a woman could be elected head). The investigators collected primarily quantitative evidence to test the effect of a women leader on public expenditures and on the gender and political attitudes and political behavior of their constituents.

The authors uncovered hugely consequential results: that in some cases women leaders increase women's political participation, that women leaders distribute public goods differently according to their own preferences (not in response to female constituents' complaints), and that exposure to a female leader weakens stereotypes about women's place in the public sphere, but that only after long term exposure does approval of women's leadership rise.

The point to take from this example and from the rest of this essay is not that qualitative measurement would have made the experimental results "richer" or more detailed, although that is certainly the case. Using qualitative research methods in this field experiment could have provided a different understanding of the causal effect, identified possible causal mechanisms of change, and framed new interpretive understandings of authority, democracy, and gender within an experimentally assessed instance of social change.

For example, participant observation of these women leaders *outside* of the council settings—for example in their homes, where they visit with other women—could have revealed whether they were influenced by women constituents in these more informal settings. Intensive interviews could have revealed how shifts in beliefs about women leaders' efficacy may have occurred—for example, did it take public statements of approval from other male council members, or elders or religious leaders? Was there a tipping point mechanism? Such qualitatively generated insights could have enabled this study to contribute more to general theories of identity, leadership, and political and social change. Moreover, ethnographic work could investigate whether understandings of authority and political legitimacy are reshaped in predictable ways by the first few woman leaders, or to what extent narratives about gender are adjusted to fit with women's new powerful role. Qualitative methods are uniquely positioned to answer these questions.

Potential problems of integrating qualitative methods with field experimentation become apparent from this brief example. For example, many qualitative methods involve more time investment and fieldwork than quantitative data collection; is the extra time feasible or worthwhile in the context of a field experiment? How could participatory or ethnographic methods measure outcomes and processes in a sufficiently large enough sample for an experimental comparison? More challenging, how can field experiments help to investigate traditionally qualitative or observational research questions about historical patterns, institutions, elites, and rare events?

In the rest of this essay, I address these concerns and expand upon some ideas about the integration of field experimentation and qualitative methods and questions. I first describe the benefits of triangulating qualitative and quantitative measurement within a field experiment, using concrete

examples from my own experimental work in Central Africa. I argue that qualitative measurement within a field experiment leads to a better understanding of the causal effect, suggests plausible causal explanations and "[extracts] new ideas at close range" (Collier 1999). I next turn to more intensive tools of qualitative inquiry, such as ethnography and interpretive work. These methods can magnify cases, social processes, and concepts within an experiment, and in some cases provide the primary data for causal inference in what Sherman and Strang (2004) term "experimental ethnography." I explore concerns about small sample sizes and scarcity of available units for random assignment. Finally, I turn to questions traditionally addressed by qualitative and observational research, including questions about historical or rare events. I propose that field experiments have a role to play in many cases, which would require disaggregating complex theories and using theory to specify a universe of cases for present-day experimental tests.

I address this essay not only to qualitative researchers, as encouragement to consider the use of field experimental methods, but also to current field experimentalists, as inspiration to adopt more qualitative approaches in their research.

Triangulating Quantitative and Qualitative Measurement Within Field Experiments

More than the Sum of their Parts

Collecting numerical, categorical, and ordinal data simplifies comparisons between experimental groups, but researchers could just as well collect and compare qualitative data from interviews, participant observation, and archives. It is widely recognized that inference is best supported by a triangulation of both types of data. Qualitative data can strengthen, modify, or altogether change the interpretation of quantitative data and describe important contemporaneous conditions of change.

The importance of triangulating quantitative evidence with qualitative evidence holds even for a great strength of field experiments—relative to laboratory experiments, field experiments are best positioned to capture effects on real world behaviors because they are located in the behavioral settings of interest (i.e., voting counties, ethnically diverse villages, credit unions). Field experiments capture behavioral data through observational techniques, but most often from public records, such as voting, public expenditure, and police records. Qualitative methods of investigation can explore what these behaviors mean in the context of the study, the possible social and political dynamics by which the behaviors were produced, ripple effects of the changes, and more.

A field experiment I conducted in eastern Democratic Republic of Congo (DRC) randomly assigned half of the radio antennae in the region to broadcast a talk show, which aired and encouraged discussion about a conflict-reduction soap opera broadcast across the entire region (Paluck 2008a). The research question was: Could the talk show increase more face-to-face discussion about conflict reduction, and would this discussion produce more favorable attitudes toward community conflict reduction techniques recommended by the soap opera? The outcome measurement, conducted with individuals in the randomly assigned regions with and without the talk show, included a close-ended survey instrument and a quantitative and qualitative behavioral measure.

In the behavioral measure, surveyors presented each study participant with a two-kilogram bag of salt at the conclusion of the survey. Surveyors told participants the salt was a thank you gift for participating in the interview, and also that a local NGO had identified a group in their community that was in need. Participants were asked if they would like to donate any portion of their salt to this group. "Which group?" participants would ask; surveyors responded per a prewritten script with, "Is there a group you would feel uncomfortable giving this to?" Nearly every participant responded by citing their "least-liked" group:[1] "Yes, the (Banyamulenge/Rega/FDLR)." To this, surveyors responded "Actually, that is the group for whom the donation drive is intended—would you still like to give?" As participants poured some amount of salt into a bag presented by the surveyor, or tied up their bag in preparation to store it away, they discussed their reasons for giving or not giving, their feelings about the donation, their expectations of the consequences of the donation, and their history with the least-liked group. The surveyors recorded this discussion as best they could by hand.

The strength of this mixed qualitative and quantitative measure was four-fold. I was able to record quantitative measures of whether and how much salt each participant gave (which I measured to the gram at the end of each day of interviews), qualitative data on the identity of each respondent's "least-liked" group, and data on participants' reasoning, feelings, and expectations about helping or not helping this group. To measure the impact of the radio talk show, I used these data and the survey responses to compare listeners in the talk show broadcast regions to listeners in the non–talk show regions.

First I coded the qualitative discussions about the salt, which were fascinatingly diverse: ranging from discussions about norms of sharing ("Congolese must pass on a gift"), to expressions of empathy and perspective taking ("When I am in great need, I know how much help from a stranger means to me"), to strategic reasons ("If I give them this salt, perhaps they will stop targeting my family"), to expressions of pure outrage ("they have killed family members, made us poor—I would rather die than help them"). This information is important and theoretically informative in its own right; I could also correlate their stated reasons and motivations with previous answers to the survey regarding their economic situation, level of education, experience during the ongoing conflict, and other variables.

These qualitative data also significantly strengthened my interpretation of the experimental effect of the radio talk show. Quantitative survey responses showed a *negative* impact of listening to the talk show—talk show listeners compared to listeners in non-talk show regions were less likely to endorse ideas for conflict reduction vis-à-vis their least liked group, and were more likely to endorse statements such as "violence is sometimes necessary in Congolese politics." The salt mea -

sure showed that talk show listeners were also significantly *less* likely to donate their salt to the needy but disliked group (74% percent of control area listeners donated salt, while 55% of talk show area listeners donated). The qualitative discussions pointed in the same (negative) direction as the quantitative survey information regarding the impact of the talk show. I further discovered that radio listeners in the talk show areas expressed significantly more outrage and grievances against the least liked group in their discussions about the salt (controlling for actual reported human rights abuses). The fact that these qualitative data were collected with a different instrument than the quantitative data strengthens support for the inference that encouraging discussion through a radio talk show had a negative impact on listeners. Even stronger triangulation would have included qualitative observations or interviews at another time or in another setting.

Causal Explanation Generation

These qualitative findings suggest a causal explanation for this negative result, specifically that talk show inspired discussion that made grievances more salient to listeners, reminding them of the hurtful actions of the other side. In general, field experimental results become exponentially more useful with these kinds of potential explanations for the process or mechanisms of change. Theory can direct a researcher's eye toward particular situations and data sources that may explain the causal chain of events, but for more exploratory research (e.g., the effect of childhood abduction into a militia on voting and political participation as an adult; Blattman 2008), deep contextual absorption ("soaking and poking" in qualitative lexicon) can inductively suggest explanations for experimentally assessed effects.

In the example of child soldiering, Blattman uses semi-structured interviews with former abducted child soldiers, leaders, and social workers to explore explanations for the finding that former child soldiers are more likely to vote. The qualitative data suggest that experience in the militia endowed former child soldiers with a sense of leadership and with a higher degree of maturity (19–20). Causal explanations suggested by such qualitative research can then be tested in successive field experiments. In my own research, I conducted the field experiment in eastern DRC because qualitative research in a previous field experiment testing conflict reduction radio soap operas suggested that discussion was an important mechanism of the observed changes in social norms and behaviors (Paluck, forthcoming). In this previous experiment, I collected systematic observations of groups listening to the treatment and comparison radio programs, and found that listeners kept up a steady rate of interjections, commentary, and side conversations during the broadcast, regarding plot developments and characters' behavior. Moreover, listeners lingered after the broadcast was over, to share their reactions and digest the messages of the show with one another. I hypothesized that face-to-face discussion about media with community members would shape perceptions of socially acceptable behavior, at least in the confines of that group. The experiment in the DRC attempted to test this causal explanation with an experiment that randomly assigned encouragement to discuss a media program via a talk show.[2]

In sum, the appeal of field experiments for qualitative researchers is that they offer the opportunity to generate strong causal inferences while "extracting new ideas at close range" (Collier 1999). I suspect that researchers who have long embraced the idea of mixed methodology will readily acknowledge this point. However, despite general enthusiasm for this idea, mixed methods have not been a common feature of field experiments.

More Intensive Qualitative Tools of Inquiry: Ethnography, Participant Observation, and Interpretive Work

More challenging than combining qualitative and quantitative data within a field experiment is integrating into an experiment qualitative methods that require intensive time investment and field engagement, such as participant observation, intensive interviews, thick description, or ethnography. This broad group of methods is often employed in the service of interpretive goals, for example complicating, historicizing, and enriching understandings of social science concepts like culture, democracy, or power (Wedeen 2002). In some cases, it may be useful to frame interpretive work within a larger field experimental test of an overarching claim of that project. Below I describe how these methods can also be used in a field experiment to investigate causal claims.

One straightforward way to integrate deep interviews, case studies, or ethnographies into an experiment is to select a reasonable number of observation units for close examination in the experimental and the control groups (see Tarrow 2004, on framing qualitative investigation within quantitative projects). Policy experiments have used this strategy—for example, the Moving to Opportunity experiment in American cities, which tested the effect of giving housing vouchers to low income residents so that they could move into better neighborhoods (Turney et al. 2006). Sociologists and anthropologists working on this project conducted repeated intensive interviews with selected men and women who were randomly assigned to receive or to wait for the voucher. The interviews explored quantitatively measured outcomes such as basic daily functioning and depression, phenomena that often require a fuller contextual understanding. In general, a feasible number of cases for intensive qualitative measurement within an experiment could be selected from each experimental group on the basis of theoretically relevant a priori characteristics to explore the contextual nature and heterogeneity of the experimentally assessed causal effects.

"Experimental Ethnography"

A more ambitious proposal in this vein is to conduct ethnographic case studies for all of the units of observation in a field experiment, or what Sherman and Stang (2004) term "experimental ethnography":

> Experimental ethnography is a tool for answering questions about why programmatic attempts to solve human problems produce what effects, on average, in the con-

> text of the strong internal validity of large-sample, randomized, controlled field experiments…This strategy can achieve experiments that create both a strong "black box" test of cause and effect and a rich distillation of how those effects happened inside that black box, person by person, case by case, and story by story (205).

Writing from the perspective of program evaluators, Sherman and Stang discuss a recent randomly implemented policy for restorative justice in England and Australia, which invited victims, perpetrators, and all those affected by the crime to meet and discuss how the perpetrator should repay his or her debt to society. When police officers offered this program to untried perpetrators and their victims, they told each party that if both parties accepted, they would have a 50% chance of having the meeting because the program was in an experimental trial.

Sherman and Stang describe how ethnographies describing the experiences of victims, perpetrators, and their families during and after the restorative justice process would have been important for fully understanding the effects of this program.[3] Specifically, experimental ethnography could use an iterative process of theory development and testing commonly associated with qualitative approaches, or grounded theory (Glaser and Straus 1967): "[t]he hypotheses that are generated from interviews or observations of one case can immediately be tested against new data on the same hypotheses collected on other cases. Even if these hypotheses and their tests are later reduced to quantitative form, the fact that they would not have emerged without ethnographic work provides a strong justification for the added cost and effort of experimental ethnography" (211).

Qualitative data on the severity of the victim's reaction to the crime, in their example, suggested the hypothesis that the magnitude of potential benefit of restorative justice on the victim's mental health was directly proportionate to the magnitude of the harm the victim suffered from the crime. The qualitative evidence both "discovered" this grounded claim and offered a way to test it, through continuous comparisons between treatment and control groups. Sherman and Stang note that it is best to conduct this kind of theory testing when *all* of the cases in an experiment can be included in an ethnography, which should be feasible for "samples of a hundred or so" (211).

Small-*N* Concerns

The Sherman and Stang proposal exposes an important tradeoff, the classic tug of war between breadth and depth that typically leaves qualitative researchers with a small sample size. Other times, qualitative researchers are restricted to a small sample size because of the limited number of units to study—e.g., only six countries that meet the criteria for a certain research topic, or 12 non-overlapping broadcast areas in the region of interest. I have two suggestions regarding this tradeoff.

Collaboration is one answer to the problem of conducting ethnography with all of the units of an experiment. Several qualitative researchers working as a team could each take responsibility for a random sample of units in the treatment and control groups. Researchers' responsibilities should overlap for a few units, as the overlapping ethnographies could serve as a continuous check on the comparability of their methods and observations. This kind of collaborative ethnography has the potential to provoke a productive discussion among ethnographers regarding the comparative versus particularistic nature of their work. The challenge of comparing their ethnographic data in the service of drawing causal inferences would require ethnographers (or intepretivists, participant observation researchers, etc.) to make their process—their definitional terms, their observational procedures, their selection of place and subjects—transparent and replicable. Such an effort would only succeed by increasing the comparative nature of the ethnographic enterprise. While some ethnographic traditions (particularly in anthropology) are opposed to the idea of producing replicable procedures and observations, this kind of a collaborative work would advance the comparative goals of researchers who are amenable to the idea.

A sustained research program is another way to accommodate a small sample in a field experiment.[4] With a small sample size, researchers may not be able to identify modest or small effects, or may over-or underestimate larger effects. My collaborator Donald Green and I have argued that in this case it is still worthwhile to do the experiment in the context of a sustained research program (Paluck and Green 2008). Repeated experiments on the same general question will average out to the true unbiased effect.

Treating Questions Typically Associated with Observational and Qualitative Investigation

One of the most frequently voiced reasons for not using field experimental methods is that a certain class of research topics are too historically based or would be unethical or impossible to test using random assignment. Questions about the historical pattern of state formation, the causes of revolutions or genocides, elite decision-making about nuclear deterrence, and the democratic peace hypothesis all fall into this category. These topics are sometimes cited as evidence that observational and qualitative researchers struggle with more "important" or "bigger" questions than those addressed by experimental methods.

Of course, field experiments (and as-if-random "natural" experiments; Dunning 2008) have already addressed many important questions that seemed unsuited to experimentation prior to their successful execution. To date, and mostly without the explicit use of qualitative methods, experiments have answered questions about the effects of political campaigns (Green & Gerber 2008; Nickerson 2008; Wanketchon 2003), police raids and crime deterrence (Sherman et al. 2002), mass media programming (Paluck forthcoming; Green and Vavreck 2008), ethnic diversity (Habyarimana et al. 2007; Posner 2004), international election monitoring (Hyde 2007), deliberative democracy techniques (Fararr et al., forthcoming; Wanketchon 2008), gender and politics (Beaman et al. 2008; J. Green 2008), corruption (Olken 2007), employment discrimination (Pager

2007), educational attainment (Sondheimer and Green 2008), health care (King et al. 2007), slavery and trust (Nunn and Wanketchon 2008), and child soldiering (Blattman and Annan 2007). Thus far, I have argued that including qualitative methods can extend the reach of field experimentalism further.

Still, causal questions rooted in history or addressing elites, violence, country-level and rare events like social movements and revolution are at one level beyond the reach of experiments. Random assignment of the purported causes of these events would be unethical or logistically impossible without dictatorial powers or a time machine. One point made in response to this dilemma is that relatively more narrow field experiments accumulate the "stubborn facts that inspire theoretical innovation" (Green 2005). Field experiments gradually collect unbiased causal facts upon which a more complex theory can be built.

I propose another idea that flows in the opposite direction. In contrast to building theories from relatively narrow empirical facts, investigators could start at the level of their highly complex theories and disaggregate them in a way that would make field experimentation possible for a few of the causal links in their specified chain. Theories of genocide, for example, make many causal claims about the road to violence. Some purported causes of genocide include elites threatened by a shift in power, bureaucratic or other tools for ethnic differentiation, land shortages, and so forth. A field experiment could not and would not randomly assign all of these conditions, but it could, for example, examine the effects of policies (introduced progressively in randomly assigned areas of the country or subsets of the population, i.e., a "random roll out") that increase or decrease ethnic differentiation (identity cards or citizenship papers), or redistribute land.

Integrating field experiments into these traditionally observational research programs in this manner would require theoretical specificity, strategic case selection (for which qualitative researchers are exceptionally qualified), and (in some cases) cooperation with policy makers or political elites. Researchers would need a high degree of theoretical specificity and clarity in that they would need to define the necessary contextual conditions of a present-day theoretical test. Some theories are intended only for historical cases (Skocpol 1977); in these instances, field experimentation would obviously reach a dead end. But for theories intended to extend into present-day contexts, researchers would need to draw out sufficient and necessary conditions for the field experimental context.

Using theories that describe necessary and sufficient conditions of the phenomenon of interest, qualitative researchers have honed the skill of case selection (Seawright and Gerring 2008) into a systematic method that requires deep contextual and historical knowledge. Selecting present-day relevant cases would be the critical task for researchers for testing theories of historical events with field experiments. Finally, many such field experimental tests would probably require collaboration with policymakers and political elites, since many of these kinds of questions involve structural, economic, or institutional shifts. Many relevant changes are occurring through new policies (again, land policies in developing countries is one example), which could be rolled out randomly. Collaborating with governments and non-governmental or international organizations presents a host of ethical and practical dilemmas, but it should not be written off as impossible. Currently, field experimentation is receiving increased respect and interest from policy makers and international organizations, mostly on the wings of the influential movement to include field experiments in development economic policy and from efforts of some political scientists as well.[5] As economists have proved with the development community, a few very useful experiments can interest stakeholders in fielding and participating in experiments of their own. Experimentation with (and even on) political elites would make the use of experimentation in observational research programs more of a possibility.

Conclusion

While I do not claim to have all of answers for how qualitative researchers can use experimental methods, I believe that researchers should not foreclose on the possibility of using field experiments in qualitative or observational research programs before considering these ideas. Integrating field experimentation into any research program will be a difficult but creative and productive process. It will require knowledge of the cases, theoretical clarity, and comparable and meaningful outcome measures. Qualitative methods, from case selection to interviewing to participatory observation, are all necessary on some level to conduct good field experiments. For this reason, qualitative researchers and current field experimentalists alike could benefit from considering the ideas I have reviewed above. Integrating qualitative methods with field experiments should encourage new and interesting investigator collaborations, and learning within all types of methodological persuasions.

Notes

[1] Based on the "content controlled" technique pioneered by Sullivan, Pierson and Marcus (1982).

[2] Note that in the DRC experiment, I was missing a critical arm of the experiment (due to logistical reasons) in which a third control group did not have access to the soap opera, which provided the topics of discussion, or to the talk show, which encouraged the discussion. Including a no-soap, no-talk control group would show (a) the effect of the soap opera, and (b) the additional effect of discussion inspired by the talk show about the soap opera. I am implementing this design in a new experiment on a peace and democracy radio campaign in Southern Sudan, by randomly assigning a radio show, discussion, radio show plus discussion, or no intervention (Paluck 2008b).

[3] They also suggest that ethnographies of the victims and perpetrators who did not accept the offer to be a part of the program would have helped explore the reach of the restorative justice program, and also more generally the ability of experimental trials to measure causal effects in a representative portion of the population.

[4] Besides the problem of low power to detect causal relationships, small samples mean that simple random assignment is more likely to create an unbalanced comparison. For example, in a sample of twelve manufacturing companies, a random "run" of similar assignment numbers might end up assigning all five car companies in the sample to the treatment condition. This problem of balance can be addressed by

matching procedures prior to randomization—simple stratification procedures in which randomization is conducted within stratified groups of car and drug manufacturing companies, for example, or more complex matching with multiple strata using statistical software (e.g., Coarse Exact Matching, Iacus et al. 2008). In my small-*N* experiments in Central Africa, I have randomized within stratified villages and broadcasting regions.

[5] EGAP, or Experiments on Governance And Politics, is one example of a recent organizational effort involving political scientists and policy organizations.

References

Beaman, Lori, Raghabendra Chattopadhyay, Rohini Pande and Petia Topalova. 2008. "Powerful Women: Does Exposure Reduce Bias?" Working paper.

Blattman, Christopher. 2008. "From Violence to Voting: War and Political Participation in Uganda." Working Paper Number 138, Center for Global Development.

Blattman, Christopher and Jeannie Annan. 2007. "The Consequences of Child Soldiering." HiCN Working Paper 22.

Brady, Henry and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Chattopadhyay, Raghabendra and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72:5, 1409–43.

Collier, David. 1999. "Data, Field Work and Extracting New Ideas at Close Range." *APSA-CP: Newsletter of the APSA Organized Section in Comparative Politics* 10:1 (Winter), 1–2, 4–6.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61:2, 282–93.

Farrar, Cynthia, James Fishkin, Donald P. Green, Christian List, Robert Luskin, and Elizabeth L. Paluck. In press. "Disaggregating Deliberation's Effects: An Experiment Within a Deliberative Poll." *British Journal of Political Science*.

Gerring, John and Rose McDermott. 2007. "Experiments and Observations: Towards a Unified Framework of Research Design." *American Journal of Political Science* 51 (July), 688–701.

Glaser, Barney and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Green, Donald P. 2005. "On Evidence-Based Political Science." *Dædalus* (Summer): 96–100.

Green, Donald P. and Lynn Vavreck. 2006. "Assessing the Turnout Effects of Rock the Vote's 2004 Television Commercials: A Randomized Field Experiment." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 20–23, 2006.

Green, Donald P. and Alan S. Gerber. 2008. *Get Out The Vote: How to Increase Voter Turnout*. Second Edition. Washington: Brookings Institution Press.

Green, Jennifer. 2008. "Mobilizing Women to Vote in Traditional Societies: An Experiment Encouraging Political Participation in Rural India." Working paper.

Habyarimana, James, Macartan Humphreys, Daniel Posner, and Jeremy Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101:4, 709–25.

Hyde, Susan. 2007. "The Observer Effect in International Politics: Evidence from a Natural Experiment." *World Politics* 60:1, 37–63.

Iacus, Stefano M., Gary King, and Giuseppe Porro, "Matching for Causal Inference Without Balance Checking." Harvard University Working Paper, available at http://gking.harvard.edu/files/abs/cem-abs.shtml.

Kahneman, Daniel, Paul Slovic, and Amos Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

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

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. "A 'Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program." *Journal of Policy Analysis and Management* 26:3, 479–506.

Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102 (February), 49–57.

Nunn, Nathan and Leonard Wantchekon. 2008. "The Trans-Atlantic Slave Trade and the Historical Origins of Mistrust in Africa: An Empirical Analysis." Working paper.

Olken, Benjamin. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (April), 200–49.

Pager, Devah. 2007. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *Annals of the American Academy of Political and Social Science* 609, 104–33.

Paluck, Elizabeth Levy. Forthcoming. "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda." *Journal of Personality and Social Psychology*.

Paluck, Elizabeth Levy. 2008a. "Is it Better Not to Talk? A Field Experiment on Talk Radio and Ethnic Relations in Eastern Democratic Republic of Congo." Working paper, Harvard University.

Paluck, Elizabeth Levy. 2008b. "A Field Experiment Testing Discussion and Media in Southern Sudan." Unpublished Paper.

Paluck, Elizabeth Levy and Donald P. Green. 2008. "Deference, Dissent, and Dispute Resolution: A Field Experiment on a Mass Media Intervention in Rwanda." Unpublished Paper.

Posner, Daniel. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98:4, 529–45.

Seawright, Jason and John Gerring. 2008. "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61:2, 294–308.

Sherman, Lawrence and Heather Strang. 2004. "Experimental Ethnography: The Marriage of Qualitative And Quantitative Research." *The Annals of the American Academy of Political and Social Sciences* 595, 204–22.

Sherman, Lawrence, Dennis Rogan, Timothy Edwards, Rachel Whipple, Dennis Schreve, Daniel Witcher, William Trimble, Robert Velke, Mark Blumberg, Anne Beatty, and Carol Bridgeforth. 2002. "Deterrent Effects of Police Raids on Crack Houses: A Randomized, Controlled Experiment." *Justice Quarterly* 12:4, 755–81.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Sondheimer, Rachel M. and Donald Green. 2008. "The Brody Paradox Revisited: Using Experiments to Estimate the Effects of Education on Voter Turnout." Unpublished paper.

Sullivan, John L., James Piereson, and George E. Marcus. 1993. *Political Tolerance and American Democracy*. Chicago: University of Chicago Press.

Tarrow, Sidney. 2004. "Bridging the Quantitative-Qualitative Divide." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 171–80.

Turney, Kristin, Susan Clampet-Lundquist, Kathryn Edin, Jeffrey R. Kling, and Greg J. Duncan. 2006. "Neighborhood effects on Barriers to Employment: Results from a Randomized Housing Mobility Experiment in Baltimore." Working paper #511, Princeton University.

Wantchekon, Leonard. 2003. "Clientalism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55 (April), 399–422.

Wantchekon, Leonard. 2008. "Expert Information, Public Deliberation, and Electoral Support for "Good" Governance: Experimental Evidence from Benin." Working paper.

Wedeen, Lisa. 2002. "Conceptualizing Culture: Possibilities for Political Science." *American Political Science Review* 96:4 (December), 713–28.

Designed-Based Inference: The Role of Qualitative Methods

Thad Dunning
Yale University
thad.dunning@yale.edu

Political scientists increasingly use natural and field experiments in their research.[1] This raises the question—how do qualitative methods contribute to these research methodologies? I suggest here that there are strong complementarities between the use of such research designs and various kinds of qualitative methods. For example, case-based knowledge is often necessary to recognize and validate a potential natural experiment. The research skills associated with qualitative fieldwork, in turn, are often required for the implementation of field experiments. Qualitative methods can be crucial for designing experimental interventions, measuring outcomes, providing evidence on mechanisms, and even constructing random assignment mechanisms.

After discussing natural experiments from a variety of perspectives, I give a short example of how a field experiment may be used to explore the relationship between cross-cutting cleavages and ethnic voting in Mali, drawing on my recent joint research on this topic. As I describe, qualitative methods have contributed in both expected and unexpected ways to this project.

Natural Experiments and Qualitative Methods[2]

An illuminating if well-known exemplar of a successful natural experiment comes from John Snow's studies of cholera transmission (Freedman 1991, 1999, 2005; Dunning 2008). While its substantive domain lies far from the concerns of most social scientists, Snow's research illustrates the key role of qualitative methods in identifying and exploiting a natural experiment to make progress on an important problem.

Nineteenth-century London suffered a number of devastating cholera outbreaks. Although predominant theories linked cholera transmission to bad air (miasma) or to ground poisons, Snow became convinced that cholera was a waste- or water-borne infectious disease (Richardson 1887: xxxiv). In Snow's research, "causal process observations" (Collier, Brady, and Seawright 2004) were crucial, both for allowing Snow to formulate a hypothesis about the causes of cholera transmission and to provide evidence for the plausibility of this hypothesis. For example, Snow noted that outbreaks seemed to follow the "great tracks of human intercourse" (Snow 1855: 2); sailors who arrived in a cholera-infested port did not become infected until they disembarked, striking a blow to the miasma theory (Snow 1855: 2).

During London's cholera outbreak of 1853–54, Snow famously drew a map showing the addresses of deceased cholera victims. Because these addresses clustered around the Broad Street water pump in the Soho district, Snow argued that contaminated water supply from the pump caused the cholera outbreak. However, there were several anomalous cases: residences located near the pump where there had been no deaths from cholera, and residences far from the pump with cholera deaths. Snow used qualitative process tracing and a heavy dose of "shoe leather" (Freedman 1991) to probe these seemingly disconfirming outcomes (Snow 1855: 39–45). At a brewery located near the Broad Street pump, where cholera death rates were anomalously low, the proprietor told Snow that a fresh-water pump was installed on the premises—and that in any case the brewers tended to drink beer, not water (Snow 1855: 42). At another address, closer to another water pump than to Broad Street—and where there had been significant deaths from cholera—Snow learned that the deceased residents had preferred, for one reason or another, to take water at the Broad Street pump (Dunning 2008). Snow's experience as a clinician, his studies of the pathology of cholera deaths, and his spot map showing the proximity of victims to the Broad Street pump all provided bits of evidence, which suggested that cholera might indeed be an infectious disease carried by waste or water.

However, Snow's most powerful piece of evidence came from a natural experiment. Large areas of London were served by two water suppliers, the Lambeth company and the Southwark and Vauxhall company. Just prior to the cholera epidemic of 1853–54, the Lambeth company moved its intake pipe further upstream on the Thames, thereby "obtaining a supply of water quite free from the sewage of London" (Snow 1855: 68), while the Southwark and Vauxhall company left its intake pipe in place. After painstaking data collection, Snow constructed a simple cross-tab showing cholera death rates during the epidemic by source of water supply. For houses served by Southwark and Vauxhall, the death rate from cholera was 315 per 10,000; for houses served by Lambeth, it was a mere 37 per 10,000 (Snow 1855, Table IX, p. 86; presented in Freedman 2005).

Why did this constitute a credible natural experiment? Unlike true experiments, the data used in natural experiments come from naturally occurring phenomena—actually, in the social sciences, from phenomena that are often the product of social and political forces. Because the manipulation of the treatment, intervention, or independent variable is not gener-

ally under the control of the analyst, natural experiments are, in fact, observational studies. However, unlike other non-experimental approaches, a researcher exploiting a natural experiment can make a credible claim that the assignment of the non-experimental subjects to treatment and control conditions is "as-if" random. Outcomes are compared across treatment and control groups, and both *a priori* reasoning and empirical evidence are used to validate the assertion of randomization.

Thus, random or as-if random of assignment to treatment and control conditions—in Snow's study, the water supply source—constitutes the defining feature of a natural experiment. This implies that at least as a necessary if not sufficient condition, the treatment and control groups are balanced with respect to other (measurable) variables that might explain cholera deaths. Notice that in a natural experiment, this is achieved not by statistical adjustment on the part of the analyst but rather by nature's as-if randomization. Snow presented various sorts of evidence to establish this pre-treatment equivalence between the two groups. In his own words,

> The mixing of the (water) supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies…It is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this" (Snow 1855: 74–75).

Moreover, residents did not appear to self-select into their source of water supply: decisions regarding water companies were often taken by absentee landlords, the decision of the Lambeth company to move its intake pipe was taken before the cholera outbreak of 1853–54, and existing scientific knowledge did not clearly link water source to cholera risk. As Snow puts it, the pipe's move meant that more than three hundred thousand people were:

> divided into two groups *without their choice, and, in most cases, without their knowledge*; one group being supplied with water containing the sewage of London, and…the other group having water quite free from such impurity (Snow 1855: 75; emphasis added).

The cholera example provides several useful lessons about the elements of a successful natural experiment (see Freedman 1991, 1999). Snow went to great lengths to gather evidence and to use *a priori* reasoning to argue that only the water supply distinguished houses in the treatment group from those in the control group, and thus the impressive difference in death rates from cholera was due to the effect of the water supply. It is also worth noting that, while the natural experiment may have been the *coup de grace* in Snow's painstaking investigation into the causes of cholera transmission, his use of this natural experiment was complemented and indeed motivated by the other evidence that he had gathered. The body of evidence Snow compiled depended on his detailed knowledge of the progress of previous cholera outbreaks in England, on his ability to cull information from a variety of sources, and especially on his willingness to do on-the-ground process tracing and close-range exploration of seemingly disconfirming cases (Dunning 2008). This kind of close-range research also gave him the information he needed to discover and exploit his natural experiment, while his apparently innate sense of good research design led him to recognize the inferential power of the approach.

Social-Scientific Examples

Several of the elements of Snow's successful natural experiment can be found in recent social-science applications, as well. Brady and McNulty (2004), for example, are interested in examining how the cost of voting affects turnout. In California's special gubernatorial recall election of 2003, in which Arnold Schwarzenegger became governor, the elections supervisor in Los Angeles County consolidated the number of district voting precincts from 5,231 (in the 2002 regular gubernatorial election) to 1,885. For many voters, the physical distance from residence to polling place was increased, relative to the 2002 election; for others, it remained the same. Those voters whose distance to the voting booth changed—and who therefore presumably had higher costs of voting, relative to the 2002 election—constituted the treatment group, while the control group voted at the same polling place in both elections.

The consolidation of polling places in the 2003 election arguably provides a natural experiment for studying how the costs of voting affect turnout. A well-defined intervention, the closing of some polling places and not others, allows for a comparison of average turnout across treatment and control groups. The key question, of course, is whether assignment of voters to polling places in the 2003 election was as-if random with respect to other characteristics that affect their disposition to vote. In particular, did the county elections supervisor close some polling places and not others in ways that were correlated with potential turnout? Brady and McNulty (2004) raise the possibility that the answer to this question is yes, and indeed they find some evidence for a small lack of pre-treatment equivalence on observed covariates such as age across groups of voters who had their polling place changed (i.e., the treatment group) and those that did not. Thus, the assumption of as-if random assignment may not completely stand up either to Brady and McNulty's careful data analysis or to *a priori* reasoning (elections supervisors, after all, may try to maximize turnout). Yet pre-treatment differences between the treatment and control groups are small, relative to the reduction in turnout associated with increased voting costs. After careful consideration of potential confounders, Brady and McNulty can convincingly argue that the costs of voting negatively influenced turnout, and a natu-

ral experimental approach plays a key role in their study.

Another increasingly common class of natural experiments exploits the existence of political or jurisdictional borders that separate similar populations of individuals, communities, firms, or other units of analysis, some exposed to a treatment or policy intervention and others not; in Dunning (2008), I review several studies and discuss the strengths and limitations of this form of natural experiments. Posner (2004), for example, studies the question of why cultural differences between the Chewa and Tumbuka ethnic groups are politically salient in Malawi but not in Zambia. Separated by an administrative boundary originally drawn by Cecil Rhodes' British South African Company and later reinforced by British colonialism, the Chewas and the Tumbukas on the Zambian side of the border are apparently identical to their counterparts in Malawi, in terms of allegedly objective cultural differences such as language, appearance, and so on. However, Posner finds very different inter-group attitudes in the two countries, with Chewas and Tumbukas in Malawi more likely to report an aversion to inter-group marriage and a disinclination to vote for members of the other group.

Posner argues convincingly that long-standing differences between Chewas and Tumbukas located on either side of the border cannot explain the very different inter-group relations in Malawi and in Zambia; a key claim is that "like many African borders, the one that separates Zambia and Malawi was drawn purely for [colonial] administrative purposes, with no attention to the distribution of groups on the ground" (Posner 2004: 530). Instead, the factors that make the cultural cleavage between Chewas and Tumbukas politically salient in Malawi but not in Zambia should presumably have something to do with exposure to a treatment (broadly conceived) on one side of the border but not on the other. Posner suggests that contrasts between inter-group attitudes of Chewas and Tumbukas in Malawi and Zambia are explained by the different sizes of these groups in each country, relative to the size of the national polities, which changes the dynamics of electoral competition and makes the groups political allies in Zambia but rivals in Malawi (see also Posner 2005).

Yet in order to argue this, Posner has to confront a key question which, in fact, sometimes confronts randomized controlled experiments as well: what, exactly, is the treatment? Or, put another way, which aspect of being in Zambia as opposed to Malawi causes the difference in political and cultural attitudes? Posner provides evidence that helps rule out the influence of electoral rules and the differential impact of missionaries on each side of the border. Rather, he suggests that in Zambia, Chewas and Tumbukus are politically mobilized as part of a coalition of Zambians living in the country's Eastern region, since alone neither group has the size to contribute a substantial support base in national elections, whereas in smaller Malawi (where each group makes up a much larger proportion of the population), Chewas are mobilized as Chewas and Tumbukus as Tumbukus (see also Posner 2005).

Clearly, the hypothesized intervention here is on a large scale—the counterfactual would involve, say, changing the size of Zambia while holding constant other factors that might affect the degree of animosity between Chewas and Tumbukus. This is quite different from imagining changing the company from whom one gets water in nineteenth-century London; one may question whether a manipulationist account of causation is most appropriate here (see Goldthorpe 2001 and Brady 2002). However, Posner's investigation of the plausibility of the relevant counterfactuals provides an example of "shoe leather" (that is, walking from house to house to find nuggets of evidence and rule out alternative explanations) in the tradition of John Snow (Freedman 1991).

In natural experiments, a key question is whether treatment assignment really is as-if random, that is, independent of other factors that might explain differences in average outcomes across treatment and control groups. The assertion of as-if random assignment may be more compelling in some contexts than in others. As I discuss in Dunning (2008), it may be useful to conceptualize a "continuum of plausibility" that assignment to treatment and control is really as-if random; in that article, I place several recent studies along such a continuum and discuss ways in which the as-if random criterion may be partially validated with evidence as well as *a priori* reasoning (Dunning 2008).

For present purposes, the central point is simply that qualitative methods and case-based knowledge may play an important role in efforts to exploit as well as to validate natural experiments. Close knowledge of specific substantive domains may allow analysts to find and exploit credible natural experiments (see also Malesky, this symposium). And while simple quantitative techniques are also important for partially validating the claim of as-if random assignment (for example, for demonstrating equivalence on measured non-treatment variables across treatment and control groups), leveraging case-based knowledge about the substantive domain under investigation is also crucial to convincing applications of the natural-experimental approach.

Field Experiments and Qualitative Methods

In a randomized controlled experiment, subjects or units are randomized to treatment and control, and the intervention or manipulation is under the control of an experimental researcher (Freedman, Pisani, and Purves 1997). The main attraction of true (randomized controlled) experiments is that they solve pervasive problems of confounding and selection bias: random assignment ensures that treated and untreated groups are equivalent prior to the intervention, up to random error.[3] With a large enough number of units, random error will play only a small role, and post-intervention differences across the treatment and control groups can be reliably attributed to the effect of treatment.

Field experiments—that is, randomized controlled experiments in which the "conditions under which a causal process of interest occurs are simulated as closely as possible" (Gerber and Green 2008)—offer many synergies with qualitative methods. As Gerber and Green (2008) point out, by definition, field experiments constitute "the conjunction of two methodological strategies, experimentation and field work." In some obvious ways, then, the skills associated with some qualitative

researchers, particularly those who do fieldwork, are requisite for field experiments as well. The close case-based knowledge associated with some qualitative research may be vital for recognizing the opportunity to conduct a field experiment, and the social and networking skills often associated with qualitative fieldwork appear to be the *sine qua non* of many field experiments, as well.

Qualitative methods may play several other important roles in field experiments, however. Although not my main focus here, one important potential contribution of qualitative methods is in identifying mechanisms, which is a crucial part of causal inference. For example, an experiment may allow the estimation of a causal effect without, however, illuminating the mechanism through which the cause produces its effect. Qualitative information may provide insights or information on context and mechanism, perhaps in the form of what Collier, Brady, and Seawright (2004) call "causal process observations." (In addition, other experiments might be designed to elucidate the mechanism).

Yet there are also many other ways in which qualitative methods can contribute to field experiments, beyond simply field research skills. For example, they can help analysts confront challenges involved in measuring outcomes, designing treatments, recruiting participants, and even randomizing subjects to treatments. My objective in the rest of this article is to describe the contributions of qualitative methods to an ongoing experiment on ethnic politics in Mali. I first describe the experiment briefly, in order to set the stage for my discussion of qualitative methods.

Cross-Cutting Cleavages and Ethnic Politics: An Experiment in Mali

Social scientists often ascribe the absence or moderation of ethnic conflict to cross-cutting cleavages—that is, the presence of alternate dimensions of identity or interest, along which members of the same ethnic group may have diverse allegiances. Despite a rich theoretical literature, however, the empirical effects of cross-cutting cleavages are notoriously difficult to estimate. One goal of my ongoing research, conducted jointly with Yale undergraduate Lauren Harrison, is to formulate an experimental method for investigating the political effects of cross-cutting cleavages.

In Mali, despite substantial ethnic diversity, levels of ethnic conflict are persistently low. Unlike some Sub-Saharan countries, parties do not form along ethnic lines, and ethnicity is a poor predictor of individual vote choice. One set of explanations advanced for this African anomaly focuses on an informal institution called *cousinage* (loosely translated as "joking cousinship"). In Mali as well as in Sénégal, the Gambia, Guinea, western Burkina Faso, and the northern Ivory Coast—areas either formerly part of the Mali Empire (c. 1230–1600) or subject since to significant immigration from those areas—families historically formed alliances on the basis of patronyms. These historical alliances are now invoked in everyday social interactions. Today in Mali, for instance, if someone with the last name Keita meets someone named Coulibaly on the street, these two fictive cousins may invoke a standard set of jokes, even if they have never previously met. The jokes reinforce the social bonds understood to inhere in their relationship.

For our purposes, these alliances constitute cross-cutting cleavages, because they occur across as well as within ethnic groups.[4] Despite a substantial literature on the alleged pacifying effects of cousinage (see Canut and Smith 2006; Davidheiser 2006: 837; Launay 2006; among early anthropologists, Mauss 1928 and Radcliffe-Brown 1940), it appears to us that this claim has not been subjected to empirical scrutiny that would allow valid inferences about causal effects. We extend the hypothesis to explain not only the absence of ethnic conflict, generically, but also the apparent absence of ethnicity in electoral politics, asking why, in an ethnically-diverse African polity, ethnicity does not predict individual vote choice, and parties do not form along ethnic lines. Our extension of the cross-cutting cleavage (cousinage) hypothesis to explain political preferences and patterns of electoral competition in Mali is new and to our knowledge has not been previously tested.

We developed an experimental design to estimate the effects of cousinage relations on evaluations of political candidates and their speeches. First, we videotaped two Malian actors delivering the same speech, which focused on standard themes in Malian political campaigns; in initial field trials in the capital of Bamako, 56% percent of experimental subjects said the speech "reminded them of a speech they had heard on a previous occasion." The speech was delivered in Bambara, which is the lingua franca of Bamako (and of Mali).[5] We then recruited experimental subjects by canvassing all of Bamako's neighborhoods (*quartiers*), approaching men and women sitting outside homes (or knocking on doors) and asking subjects if they would participate in a study on political speeches.[6] We administered a screening questionnaire to each potential subject, asking for each subjects' first and last name and ethnic identity, along with various other personal information; this allowed us to assign subjects randomly to the treatment conditions, as described below.[7] Experimental subjects then viewed our videotaped political speeches on a portable DVD player or laptop, using headphones.[8] Finally, subjects then answered questions about the content of the speech and the politician who delivered it. For instance, they answered questions about the global quality of the speech, whether the speech made them want to vote for the candidate, and specific questions about candidate attributes such as competence, likeability, and intelligence.

The manipulation in this experiment consisted of what subjects were told about the politician's last name. In Mali, last name conveys information about both ethnic identity and about cousinage ties. Thus, varying the politician's last name allowed us to vary the treatment along two dimensions: the ethnic relationship of the politician and the subject (same ethnicity/different ethnicity) and their cousinage relationship (joking cousins/not joking cousins). Our resulting experimental design had six treatment conditions, four of which are shown in the cells of Table 1. We also added a fifth condition, in which the subject was provided with no information about the last name of the politician (and thus no information about

ethnicity or cousinage ties), and a sixth treatment condition, in which the politician had the same last name as the subject.[9]

According to our hypotheses, a joking cousin relationship between voters and politicians should moderate the negative effect of ethnicity on voters' evaluations of politicians. We expect evaluations of politicians to be more positive on average if the politician is a co-ethnic: thus, in Table 1, we expect to find that mean evaluations of co-ethnic politicians (first row) are more positive than mean evaluations of non co-ethnics (second row). On the other hand, we also expect joking cousins to be evaluated more positively than non-joking cousins, so that mean evaluations of subjects in the first column are more positive than evaluations in the second column. The main point, however, is that we expect non-coethnic cousins (top-right cell) to be evaluated more positively than non-coethnic non-cousins (bottom-right cell).[10] Such a finding would be consistent with the idea that due to cousinage relations, members of the same ethnic group have diverse allegiances along a cross-cutting dimension of identity.[11]

Table 1: Experimental Design (Four of Six Treatments)

	Joking cousins	Not joking cousins
Same Ethnicity		
Different Ethnicity		

We began rolling out this experiment at the end of July 2008; though we have finished initial field-testing at the time of writing, we have not yet seen data from the main phase of data collection. The publication of hypotheses in this newsletter constitutes a public posting of the experimental protocol prior to analysis of the data. Our principal form of analysis for testing these hypotheses will be difference-of-means tests across subjects randomly assigned to each of the six treatment conditions, with ancillary testing of sub-groups due to our interest in possible treatment effect heterogeneity.

In the interest of brevity, I will now describe just two areas in which qualitative methods have been crucial in designing and implementing this experiment: the design of the experimental stimulus, and the creation of a cousinage matrix that allowed us to assign subjects to treatment conditions.

The Experimental Stimulus: Writing a Typical Political Speech

Our goal in designing the experimental stimulus was to create a speech that would engage subjects' attention while mimicking as closely as possible a typical political speech given by a candidate for deputy in the legislature. Here, one of us (Lauren Harrison) drew on earlier fieldwork in which she observed parliamentary campaigns in Bamako in 2007. After comparing our speech to transcripts of real political speeches, we vetted the speech with several Malian informants. I will not belabor the point here but will simply point out that fieldwork and other qualitative methods played an important role in the design of the experimental treatment.

Random Assignment: Creating a Cousinage Matrix

More involved fieldwork was required for the second topic I will discuss here. In order to assign subjects at random to one of the six treatment conditions, we created a large matrix, each row of which corresponds to a Malian last name that we could expect to encounter in the field.

For instance, Table 2 shows a row of the matrix for a person named Keita from the Malinké/Maninka ethnic group. The columns of this row give the last names associated with each of our six treatment conditions. For example, the names in the first two columns are all from the same ethnic group, but Sissoko and Konaté (first column) are considered cousins of the Keita, while Diané (second column) is not. The names in the third and fourth columns, on the other hand, are names associated with other ethnic groups, some of them cousins of the Keita (third column) and some of them not (fourth column). Note that in cells with multiple entries, such as in the first, third, and fourth column in Table 2, the politician's assigned last name was selected at random from the names in the cell.

Table 2: A Typical Row of our Random Assignment Matrix

(1) Co-ethnic/ Cousin	(2) Co-ethnic/ Not cousin	(3) Not co-ethnic/ Cousin	(4) Not co-ethnic/ Not cousin	(5) No Name	(6) Same Name
Keita (Maninka)	1. Sissoko 2. Konaté	1. Diané	1. Doucouré 2. Sacko 3. Sylla 4. Coulibaly 5. Touré	1. Diallo 2. Cissé 3. Dambelé 4. Théra 5. Dabo 6. Togola 7. Watarra	Pas de nom	**Keita** (Maninka)

Qualitative fieldwork was crucial for constructing this cousinage matrix. Before arriving in Bamako, we reviewed the secondary literature and conducted interviews with experts on cousinage as well as ordinary Malian informants. This enabled us to determine, as an initial matter, the cousins that are associated with many Malian last names and to construct a preliminary, skeletal matrix. Upon arrival in Mali, we solicited feedback on the matrix from key informants and, with their help, added to the list of names included in the left column (that is, the names of potential subjects) and also refined the list of politicians' names included in each column of each row.

Next, we field-tested an initial version of the matrix on 169 subjects. Data from this initial field trial, as well as additional qualitative information obtained in the field, allowed us to expand and improve the matrix again, and 47 more subjects participated in a second phase of the experiment using our improved matrix. Finally, in mid-August 2008, we revised the matrix once again, for reasons discussed below; this final revised matrix is being used to roll out the experiment during September 2008. Our final version of the matrix includes more than 200 names in the left-hand column, including all of the most typical Malian names.

In our initial field trials, experimental subjects did not always perceive themselves to be in the correct cell—that is, the treatment condition to which they had been randomly assigned. In fact, subjects inferred ethnicity with great accuracy: given only the last name of the politician, and choosing from more than 14 possible ethnic categories, subjects correctly classified the politician's ethnicity 75% of the time. However, in initial trials, they more frequently labeled cousins as non-cousins, or non-cousins as cousins.

This mismatch in initial trials between the treatment conditions to which some subjects were assigned and the treatment conditions they perceived, raises important inferential issues.[12] After all, what we care about in this study is the effect of subject perceptions—we want to know how *perceiving* oneself as being a cousin or not being a cousin of the politician, or his co-ethnic or not, shapes evaluations of the candidate's speech. Here, the mismatch probably occurred for two reasons. First, correctly classifying cousinage relations for over 200 last names is difficult; our initial matrix of cousinage relations was highly imperfect. In this experiment, there was a tradeoff involved in limiting the names of potential subjects. On the one hand, cousinage relations are much better understood by us (and by Malians) for a few very common names, such as Keita, Coulibaly, Touré, or Cissé, than for less common names, so we might have had a better overall accuracy/compliance rate had we limited the study population to subjects with such last names. On the other hand, limiting the number of names would have meant more inefficient and costly subject recruitment.

Second, however, even if we could create a perfectly accurate matrix of cousinage relations, as understood by key informants, people vary in their knowledge of cousinage relations in Mali. For instance, are the Keita and the Doucouré (third column of Table 2) really cousins? Reasonable minds can apparently disagree. As one leading expert on cousinage puts it, "The question of which *jamu* [patronym] actually jokes with whom is subject to considerable indeterminacy. Lists of the joking partners of any given *jamu* may vary from community to community, or even from individual speaker to speaker" (Launay 2006: 799). Our own experience in the field validated this observation.

The key to resolving this conundrum is that some cousinage links are in fact widely understood: everyone agrees that the Keita and the Coulibaly are cousins. We therefore took the approach of limiting names in the first and third column of Table 2 to those *vrai cousins* or true *senanku* (the Bambara word for cousin), while also only including names in the second and fourth cell that we thought would maximize the chance of correct identification as non-cousins. We devoted considerable effort in the field to accomplishing this task, with the help of key informants. Initial indications suggest that our revised cousinage matrix is allowing much greater accuracy in subject assignment to treatment during the main roll-out of the experiment.

The point is that eliciting a reliable map of cousinage relations from key informants very centrally involved qualitative as well as mixed methods. For instance, to revise our cousinage matrix we conducted qualitative interviews with key informants. We then also employed quantitative analysis of the experimental data from initial trials. To improve the cousinage matrix, we therefore iterated between focused interviews, new versions of the cousinage matrix, and our experimental data to improve the random assignment mechanism in this experiment.

Finally, qualitative methods will likely play a key role in interpreting the results of the experiment—for example, in assessing the extent to which the experimental results can allow us to infer that cousinage plays the political role attributed to it. Here, we will want to analyze the potential role of cousinage in important parliamentary and presidential electoral campaigns.

Conclusion

Natural and field experiments are assuming a place of greater prominence in political science. They also appear to offer substantial opportunities to qualitative researchers. The type of experiment I described in Mali can be implemented relatively inexpensively; in fact, such a project would probably be well within reach for a graduate student working on his or her dissertation. Most importantly for present purposes, natural and field experiments often require skills and case-based knowledge associated with qualitative research. The inferential advantages of natural and field experiments may be increasingly combined with the strengths of qualitative research to generate new forms of mixed-method research, in the service of research programs in many different substantive areas.

Notes

[1] For evidence on the growing use of field and natural experiments, see Green (2007), Gerber and Green (2008), or Dunning (2008).

[2] Some of the material in this section is based on Dunning (2008);

I am grateful to *Political Research Quarterly* and to co-editor Amy Mazur for permission to use the material.

[3] Of course, problems of post-intervention bias can arise: subjects who get the vaccine may tend to go swimming.

[4] For example, the Keita are part of the Malinké ethnic group, while their joking cousins the Coulibaly are part of the Bambara ethnic group.

[5] Though Bambara is the first language of one ethnic group in Mali, its use does not imply a particular ethnic identity on the part of the politician. When experimental subjects were not provided with the politician's last name, their guesses about his ethnicity closely tracked the distribution of ethnic groups in Bamako.

[6] The experimental population is a convenience sample, but distributions on several measured variables are similar to those given by the census for Bamako. However, the experiment under-represents women.

[7] First name and other identifying information of subjects were subsequently discarded, as described in our protocol approved by Yale's human subjects review board.

[8] Only experimental subjects could hear the speech through the headphones, and only one subject was recruited from any group; subjects also answered follow-up questions on their own. This limited the potential that subjects' responses to treatment depended on the treatment assignment of other subjects.

[9] The sixth treatment may allow us to distinguish a "same ethnicity" or a "joking cousin" effect from a mere "sameness" effect: perhaps people simply want to vote for politicians who share their last names.

[10] However, based on our qualitative research, we believe that subjects may not clearly distinguish between cousins and non-cousins, among their co-ethnics.

[11] We do not have strong expectations about the sign of any interaction between co-ethnicity and cousinage.

[12] From an experimental design perspective, this issue can be analogized to the problem of compliance with an experimental protocol. See Freedman (2006) for a discussion of relevant analytic approaches.

References

Arceneaux, Kevin, Donald Green and Alan Gerber. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 37–62.

Berk, Richard, and David Freedman. 2008. "On Weighting Regressions by Propensity Scores." *Evaluation Review* 32: 392–409. Available online at http://www.stat.berkeley.edu/~census/weight.pdf.

Brady, Henry E. 2002. "Models for Causal Inference: Going Beyond the Neyman-Rubin Holland Theory." Presented at the Annual Meeting of the APSA Political Methodology Working Group, Seattle, Washington, July 16.

Brady, Henry E. and John McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Presented at the Annual Meeting of the Society for Political Methodology, Stanford University, July 29–31, 2004.

Canut, Cécile, and Étienne Smith. 2006. "Pactes, Alliances et Plaisanteries. Pratiques Locales, Discours Global." In Cécile Canut and Étienne Smith, eds. "Parentés, Plaisanteries et Politique," special issue of *Cahiers D'Études Africaines* XLVI:4, 795–808.

Collier, David, Henry E. Brady and Jason Seawright. 2004. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanaham, MD: Rowman & Littlefield.

Davidheiser, Mark. 2006. "Joking for Peace: Social Organization, Tradition, and Change in Gambian Conflict Management." In Cécile Canut and Étienne Smith, eds. "Parentés, plaisanteries et politique," special issue of *Cahiers D'Études Africaines* XLVI:4, 795–808.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching:A Reply to Smith and Todd." *Journal of Econometrics* 125:1, 355–64.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94: 1053–62.

Doherty, Daniel, Donald Green and Alan Gerber. 2006. "Personal Income and Attitudes toward Redistribution: A Study of Lottery Winners." *Political Psychology* 27:3.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61:2, 282–93.

Dunning, Thad, and Susan Hyde. 2008. "The Analysis of Experimental Data: Comparing Techniques." Presented at the Annual Meeting of the American Political Science Association, Boston, MA, August 27–September 2.

Freedman, David. 1991. "Statistical Models and Shoe Leather." In P.V. Marsden, ed., *Sociological Methodology* 21. Washington, DC: American Sociological Association.

Freedman, David. 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14: 243–58.

Freedman, David. 2005. *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press.

Freedman, David. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691–713.

Freedman, David, Robert Pisani, and Roger Purves. 1997. *Statistics*. Third ed. New York: W.W. Norton, Inc.

Gerber, Alan S. and Donald P. Green. 2008. "Field Experiments and Natural Experiments." *Handbook of Political Methodology* (forthcoming).

Goldthorpe, John. 2001. "Causation, Statistics, and Sociology." *European Sociological Review* 17:1, 1–20.

Green, Donald P. 2007. "Experimental Design." *Encyclopedia of Research Methods in the Social Sciences* (forthcoming).

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth-Century Retrospective." *Quarterly Journal of Economics* 115:1, 45–97.

Launay, Robert. 2006. "Practical Joking." In Cécile Canut and Étienne Smith, eds. "Parentés, Plaisanteries et Politique," special issue of *Cahiers D'Études Africaines* XLVI:4, 795–808.

Mauss, Marcel. 1928. "Parentés à Plaisanterie." Annuaire de l'École pratique des hautes études, section des sciences religieuses ("Les classiques en sciences sociales") Melun, Imprimerie administrative, Paris: 3–21.

Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98:4, 529–45.

Radcliffe-Brown, A. R. 1940. "On Joking Relationships." *Africa* 19: 133–40.

Snow, John. 1855. *On the Mode of Communication of Cholera.* London: Churchill. Reprinted in *Snow on Cholera*, London: Humphrey Milford: Oxford University Press, 1936.