The Rise of Experimentation in Political Science

RONALD ROGOWSKI

Abstract

Experimental research has expanded markedly in political science over the past 30 years: the number of experimental articles in the *American Political Science Review* has almost quintupled since the mid-1980s. The main reason is intellectual: most scholars by now agree that random assignment of cases to "treatment" provides the most (perhaps the only) convincing evidence of causation. The second reason is technical advances that permit kinds of experimentation that, before about 2000, hardly existed: field, natural, and survey experiments. These have grown, while laboratory experiments have receded. While concerns remain about the external validity of these experiments, both journals and funding agencies will likely move increasingly in this direction.

Three reasons may be advanced for the surge in experimental studies in political science since the turn of the millennium: (i) technological advance; (ii) demands for greater rigor and replicability from funding agencies, peer reviewers, and informed publics; and (iii) most importantly, the greater ability of experiments—some, assuredly not all—to prove what causes political phenomena.

That there has been a huge rise is hardly in dispute. The share of articles in leading journals that report on results of some kind of experiment has risen sharply (Figure 1); articles that themselves report on the rise have multiplied (Druckman, Green, Kuklinski, & Lupia, 2006; McDermott, 2002); and not only handbooks (notably Druckman, Green, Kuklinski, & Lupia, 2011) and how-to manuals on experimentation (Gerber & Green, 2012) have appeared but, starting in 2014, a new journal devoted entirely to experimental work.[1]

Less evident has been a shift toward new and different kinds of experimentation. The once-dominant laboratory experiments, conducted chiefly among undergraduate subjects at major universities, have yielded to *survey*,

1. *Journal of Experimental Political Science*, eds. Rebecca B. Morton and Joshua A. Tucker. Cambridge Journals.

Emerging Trends in the Social and Behavioral Sciences.
Robert Scott and Marlis Buchmann (General Editors) with Stephen Kosslyn (Consulting Editor).
 ISBN 978-1-118-90077-2.

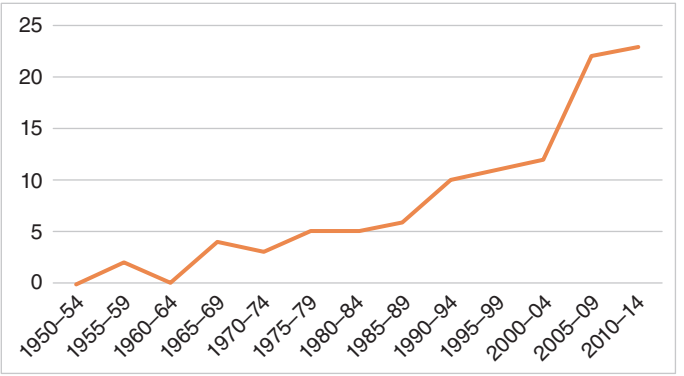

Figure 1 Experimental Articles in APSR, 1950–2014. (Note: Between 1906 and 1954, no experimental articles appeared.)

field, and "natural" experiments. In each case, the Holy Grail is fully randomized "treatment" of some presumably representative set of subjects. If, for example, a random subset of voters is allowed to register more conveniently, by how much, if at all, does the turnout of that "treated" subset change? The virtue of random assignment, no less in politics than in medicine, is that it presumably controls for all other sources of variance in the outcome. If assignment is truly random, then the members of the "control" and the "treatment" group can differ hardly at all in their composition by gender, age, political inclination, previous voting history, or indeed any other characteristic. They should of course also not differ in their awareness of whether they have been "treated," so most experiments take care to hew to a "single blind" standard, in which such knowledge is withheld from the subjects; few, so far, have also attained the medical "gold standard" of double-blindedness (in which the researchers, too, remain ignorant of who has, and has not, been treated until the study is complete).

Survey experiments best exemplify the effects of newer technology and most readily achieve true randomness. Services such as YouGov, now available in 37 countries (https://yougov.co.uk/about/our-panel/), offer cheap and ready access to nationally representative panels of Internet responders, who can then be randomly divided into various treatment and control groups. In the pioneering study by Michael Tomz and Jessica Weeks (2013), for example, respondents in the United Kingdom and the United States were presented with an identical scenario—a hypothetical country is surreptitiously developing a nuclear weapon—and asked whether a preemptive strike was justified. In half the sample, however, the fictitious country was described as a democracy, and in the other half as an autocracy. In both countries, respondents were far likelier to endorse a preemptive attack against the autocracy,

thus suggesting a potent reason for the prevalence of the "democratic peace," that is, the rarity of armed conflict between democratic states.

What this example also demonstrates, however, along with many others, is the well-known problem of "external validity," is the Internet panel fully representative of the citizenry, would subjects respond in the same way to an actual crisis as to the hypothetical one, how easily could leaders persuade them that a threatening democratic state was actually an autocracy, and so on? We already have some evidence, albeit from a less fraught issue area, that people's responses to hypothetical choices fail to predict how they will behave when faced with the same real choices, with real consequences (Barabas & Jerit, 2010).

Field experiments face far fewer problems of external validity, approximating as they do the "silver standard" of randomized (but by no means double-blind) field tests (RFTs; cf. Manzi, 2012, p. 77). In the typical case, as indeed in what is usually credited as the very first experiment in political science (Gosnell, 1926), voters, legislators, voting districts, or even the coverage areas of television stations (Gerber, Gimpel, Green, & Shaw, 2011) are randomly assigned some treatment—a get-out-the vote mailing, a mailing that invokes peer pressure by listing which of their neighbors recently voted, a specific television advertisement, the same advertisement in a different but widely spoken language (usually Spanish)—while others are assigned a different treatment or no treatment at all. The differences in response constitute powerful evidence both of the effect (or lack thereof) and of its magnitude.

Even here, doubts about external validity may arise: would voters in, say, Kansas respond in the same way as voters have been demonstrated to do in, for example, Connecticut or Texas? However, the greater obstacles to such experiments are ones of expense and nondeception. Mailings are expensive, door-to-door campaigns even more so; ethical researchers cannot run ads or send mailers that falsely imply they come from a given candidate, political party, or official agency; nor can the treatment involve untruths (e.g., attributing to a candidate something she never said or a political affiliation he does not have).

Scholars have proven remarkably adept at avoiding these pitfalls. In "endorsement" surveying, for example, one can state a position that opposing sides have in fact taken (Taliban vs ASIF; Republicans vs Democrats) but in half of the sample attribute it solely to one group (e.g., Taliban), in the other half to their opponents (in this case, ASIF) and in each case ask respondents whether they agree with that position (Lyall, Blair, & Imai, 2013). If support differs markedly according to which side is named as having "endorsed" the position, that may be taken as tacit support of the one side or the other. On even more delicate topics, or ones where respondents may fear that

their answers will elicit retribution, "list" or "noise-introducing" techniques may be used (Blair, 2015; Blair, Imai, & Lyall, 2014).[2] Equally inventive, and involving even more delicate issues of funding and of scholarly detachment, was a pioneering study in which researchers persuaded a primary election candidate for statewide office in Texas to permit them to assign his different television ads, in English and in Spanish, randomly to different metropolitan areas (Gerber *et al*., 2011). Because the campaign was also funding rolling surveys, it proved possible not only to measure the impact of a given advertisement but the duration of the impact. (For attack ads, the effect waned quickly.)

Some of the most daring and original field experiments have been performed in poorer and less stable countries, even in ones that have experienced recent civil war or genocide. In Rwanda, for example, Paluck and Green (2009) ascertained, again by random assignment of subjects in a small set of rural villages that had experienced the genocide, whether regular viewing of a government-sponsored *telenovela* that subtly advocated ethnic reconciliation (vs a "control" set of videos on AIDS prevention) actually changed expressed attitudes and—far more difficult to pin down—behaviors.[3]

Finally, and most controversially, *natural experiments* have proliferated. In these, there is no attempt to assign subjects or regions randomly; rather, evidence is advanced that some exogenous event—the building of an express highway (Shami, 2012), the drawing or redrawing of state, regional, or electoral district boundaries (Ansolabehere, Snyder, & Stewart, 2000), which party held power in London at the time a given Indian province was subjected to British rule (Banerjee & Iyer, 2005), whether military recruits from a given area experienced combat (Jha & Wilkinson, 2012)—has yielded a "quasi-random" assignment of units or persons to "treatment" and "control" groups. An important subset of natural experiments (although not always so regarded) involves so-called regression discontinuity analyses, in which cases on slightly to one side of a "discontinuity" are compared with those on the other side. In one of the most common examples of such studies, the researcher compares parliamentary candidates who narrowly won their

2. These techniques obscure the responses of individuals—in that sense, they introduce "noise" into the "signals" from individual respondents—but still permit valid inference from the set of respondents as a whole. In the "list" technique, respondents are asked only to indicate *how many* of the items on a larger list they endorse, not which ones. In the survey in Afghanistan, for example, the question asked was, "I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support" (Blair *et al*., 2014, p. 1045).

3. Pushed too far, such field experiments raise ethical concerns: Are subjects, for example, being deceived or being surreptitiously observed? A different concern is that some less developed countries can become such frequent loci of field experiments that subjects become too knowledgeable, or even jaded.

seats with rival candidates who narrowly lost. Here the assumption is maintained that such narrow victories and defeats—whether a candidate receives 49.9% or 50.1% of the vote—are essentially decided by chance, so that in all other respects these narrow winners are identical to the narrow losers. [See, for one recent example, Eggers and Hainmueller (2009).]

Interestingly, one of the earliest known practitioners of the natural experiment was Abraham Lincoln. In the debates in the 1850s over whether to permit slavery to expand into the territories, Lincoln's adversary Stephen Douglas argued that the legal status of slavery did not matter: slavery flourished where soil and climate encouraged plantation agriculture, withered where they did not. In rebuttal, Lincoln in his famous "Cincinnati Speech" of September 17, 1859 (Lincoln, 1897) offered a persuasive natural experiment:

> Let us take an illustration between the States of Ohio and Kentucky. Kentucky is separated by this River Ohio, not a mile wide. A portion of Kentucky, by reason of the course of the Ohio, is further north than this portion of Ohio, in which we now stand. Kentucky is entirely covered with slavery; Ohio is entirely free from it. What made that difference? Was it climate? No! A portion of Kentucky was further north than this portion of Ohio. Was it soil? No! There is nothing in the soil of the one more favorable to slave labor than the other. It was not climate or soil that caused one side of the line to be entirely covered with slavery, and the other side free of it. What was it? Study over it. Tell us, if you can, in all the range of conjecture, if there be anything you can conceive of that made that difference, other than that there was no law of any sort keeping it out of Kentucky, while the Ordinance of '87 kept it out of Ohio.

Yet, as Lincoln was wise enough to note, the presumption of "quasi-randomness" remains always rebuttable in natural experiments. There may indeed be some difference in "all the range of conjecture" that determines or explains the contrast between the supposed treatment and control groups, or that shows them to differ substantially from each other in some way that random assignment could not yield. If, in a regression discontinuity analysis, three-quarters of the narrow victories in Congressional races go to Republicans, while three-quarters of the narrow losers are Democrats, the quasi-random assumption is violated in an important way. If the systems of land tenure imposed by the British in India resulted more from preexisting post-Mughal institutions than from the colonial masters' prevailing ideology (Foa, 2015), we can no longer assume random assignment. Had it been the case—and the authors take care to show that it was not—that combat experience of Indian soldiers in World War II had been confined largely to groups that the British had regarded as "warlike castes," again the assignment to treatment or control would have been far from random.

Trenchant criticisms of some early natural experiments along these lines have been raised, most pithily in an article aptly entitled, "When 'Natural Experiments' Are Neither Natural nor Experiments" [Sekhon and Titiunik (2012); but see also Imai, Keele, Tingley, and Yamamoto (2011)]. Such exposés by no means rule out natural experiments—Lincoln's remains persuasive to the present day—but they do raise a warning flag: one must demonstrate convincingly that a "quasi" (literally: "as if") random assignment has a high likelihood of being close to random.

In a few extraordinarily lucky instances, a natural experiment offers full and undoubted randomization. Robert F. Erikson and Laura Stoker (2011), for example, exploited the Vietnam-era draft lottery, in which men's birthdays were literally chosen by being drawn from a rotating drum, to establish that having been assigned a low "draft number" and, hence exposed to a risk of conscription, was strongly and enduringly associated with more pacifist and leftist political attitudes.

Finally, *laboratory experiments* continue to be performed, but they seem to be fading from the mainstream journals. From 2012 to the present (Autumn 2015), for example, only two articles based on laboratory experiments have appeared in the *APSR*, about 10% of the total number of experimental papers published during that time. What seems to have undermined this kind of experimental work is a growing skepticism about its external validity: do ordinary people, or actual decision-makers, behave in the same way as undergraduate subjects sitting before a computer? In some cases, practitioners of this approach have actually tried to buttress their findings by replicating their studies among more representative subjects. Laboratory experiments probably convince best when they tap basic human characteristics—our propensity to miscalculate odds, or to draw false inferences by "thinking fast" (Kahneman, 2011)—rather than specifically political reactions. In other words, it seems that laboratory experiments are migrating to the domain of psychology and away from political science.

The trend therefore is clear; but what has driven it? The first factor, as mentioned earlier, is simply improved technology. By a corollary of "Moore's Law," computerization has made all kinds of experiments cheaper and easier to do, albeit by varying margins. Survey experiments would have been impossible, or at the very least prohibitively expensive, in the precomputer age. Field experiments have been aided by video technology, smart phones (e.g., to register responses instantly or to photograph electoral irregularities), and faster communications from the field to a central data-gathering point. Even laboratory experiments have benefited significantly: subjects who once registered their responses laboriously, often on paper, now do so instantly from a computer screen. In addition, natural experiments are more easily elicited through such innovations as digitized archives and GIS mapping.

A second factor, often underestimated, has involved demands from politicians, opinion makers, and even informed publics for clearer, more convincing, and more reliable findings, ones on which policymakers can actually rely. The critic Jim Manzi (2012), whose work the *New York Times* columnist David Brooks (2012) popularized, contends that in the most important recent crises—the Great Recession, the Iranian nuclear threat, rising income inequality, overtime variations in crime rates—the social sciences, including economics, have relied on models that, while "useful" and "interesting," do "not establish a causal relationship with sufficient certainty to [permit] rational prediction of the effect of a change in policy (Manzi, 2012, p. 105)." In his view, only fully randomized testing can establish causation with sufficient certainty; and such testing must set itself more modest goals, elucidating pieces of the puzzle—as, in fact, most recent experiments do.

While much of the rising political opposition to governmental funding of the social sciences betrays simply a "shoot the messenger" attitude toward findings the politicians or opinion leaders dislike (cf. Rogowski, 2013), political science in particular renders itself vulnerable to attack on precisely the ground Manzi and Brooks suggest: weak and tentative causal inference, occasioned entirely by lack of fully randomized testing. To draw an unhappy parallel, if medicine still relied on anecdote and induction, the National Institutes of Health might be similarly subject to political attack. Even massively evidence-supported science, of course, can be attacked by the ignorant, the deluded, or the meretricious[4]—think only of the current controversies over global warming, vaccines, and genetically modified organisms—but refuting such nonsense is far easier if the truth is supported by extensive and repeated randomized tests, as for example in the case of the vaccine controversy.

Peer reviewers for major journals in political science, to judge by the experience of the *APSR* and other leading outlets, have begun, if not to reject other modes of inquiry, to credit far more the findings of experimental research. More specifically, the now-frequent criticism of endogeneity, that is, of possible reverse causation or of causation of two correlated variables from a third factor not considered, has felled many an otherwise convincing paper;[5] and the by-now conventional answer of employing a clearly exogenous instrumental variable (IV) can never be as convincing as a randomized experiment—although, of course, such experimentation is hardly possible in the case of historical data.

4. Thus tobacco companies massively funded research that cast doubt on the link (which they knew perfectly well to be irrefutable) between smoking and lung cancer; and, in more recent years, large oil companies have sought to create controversy over man-made global warming, while in private they not only accepted its reality but were taking measures to protect their own investments against such consequences as rising ocean levels (Krauss, 2015; Lieberman & Rust, 2015).

5. I speak here, albeit only impressionistically, from my 4-year experience (2008–2012) as lead editor of the *APSR*.

The National Science Foundation (NSF) is moving only glacially in this direction, likely to its political detriment. Of 237 NSF political science awards with start dates in the 4 years between September 1, 2011 and August 31, 2015, 16, or not quite 7%, included the words "experiment" or "experimental" in their titles, and the trend appeared virtually flat: 6% in the first 2 years, just under 8% in the last two.[6] Yet as it becomes evident that the few experimental projects they have funded have been among those with highest impact, and as it becomes more urgent to fend off political attacks, one suspects that a greater share of funding will go to experimental work. For now, however, some of the most startling and fine-grained studies, including notably the one from the Texas primary election, are being funded by campaigns, other private sources, or smaller foundations.

The parallel shift in economics, albeit more recent, seems also to be more rapid, particularly in development economics, where the work of Abhijit Banerjee and Esther Duflo (2011) has proved revolutionary and is rapidly coming to dominate. The World Bank, the Inter-American Development Bank, and the governments of Indonesia and India, among many others, have tested policy innovations through randomized controlled trials (RCTs) (*Economist*, 2013, 2015).

The final advantage, and certainly the most important one, is sheer scientific beauty and honesty. We understand almost immediately that a fully randomized experiment convinces us as no nonexperimental method can. At their best, the findings of randomized experiments in political science are as exciting and important as those of the earliest ones in medicine—which, surprisingly, date only from the late 1930s (Manzi, 2012, Chapter 7).

It should go without saying that many important questions in political science, especially ones of historical causation, cannot be addressed by randomized trial—although even in cases from the past, there is great value in searching, as Lincoln did and Banerjee and Iyer have done, for some natural experiment. Other large questions will have to be addressed piecemeal: Tomz and Weeks have by no means solved the whole puzzle of the democratic peace, but they have certainly shed important light on how much of it is due, precisely as Immanuel Kant originally argued, to citizens' reluctance, on both moral and practical grounds, to enter combat with another democracy. Furthermore, we do not know whether, in general, exposing citizens to the risks of combat by conscription makes them in later years more bellicose or more pacific (or neither); but thanks to Erikson and Stokes, we now know for certain that, in the case of the United States during the Vietnam war, it made potential draftees lastingly more pacific.

6. Search on NSF Awards database (http://www.nsf.gov/awardsearch/), using "advanced search" function and keying for SBE Directorate and Program "political science," within the specified dates. Search conducted on October 16, 2015.

It requires little courage to make bold predictions about a future in which one will no longer be alive, but I will at least speculate that, 50 years from now, experimentation will play much the same role in political science as it now does in medicine, and with similar results in public confidence and material support. If I am wrong, and if political science remains a largely nonexperimental science, it will continue to be treated with skepticism and stinginess—and, I fear, rightly so.

REFERENCES

Ansolabehere, S., Snyder, J. M., Jr., & Stewart, C., III, (2000). Old voters, new voters, and the personal vote. *American Journal of Political Science*, *44*, 17–34.

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, NY: PublicAffairs.

Banerjee, A., & Iyer, L. (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *American Economic Review*, *95*, 1190–1213.

Barabas, J., & Jerit, J. (2010). Are survey experiments externally valid? *American Political Science Review*, *104*, 226–242.

Blair, G. (2015). Survey methods for sensitive topics. *APSA Comparative Politics Newsletter.*

Blair, G., Imai, K., & Lyall, J. (2014). Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science*, *58*, 1043–1063.

Brooks, D. (2012). Is our children learning? *The New York Times*, 26 April.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2006). The growth and development of experimental research in political science. *American Political Science Review*, *100*, 627–635.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (Eds.) (2011). *Cambridge handbook of experimental political science*. Cambridge, England: Cambridge University Press.

Economist. (2013). Random harvest: Once treated with scorn, randomised control trials [sic] are coming of age. 14 December.

Economist. (2015). Randomized controlled trials: measure for measure. 12 December.

Eggers, A. C., & Hainmueller, J. (2009). MPs for sale? Returns to office in postwar British politics. *American Political Science Review*, *103*, 1–21.

Erikson, R. F., & Stoker, L. (2011). Caught in the draft: The effects of Vietnam draft lottery status on political attitudes. *American Political Science Review*, *105*, 221–237.

Foa, R. (2015). *The pre-colonial origins of state capacity: Evidence from Indian districts*. Delivered at annual meeting of the American political science association, San Francisco.

Gerber, A., Gimpel, J. G., Green, D. P., & Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *American Political Science Review*, *105*, 135–150.

Gerber, A., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York, NY: W. W. Norton & Co.

Gosnell, H. F. (1926). An experiment in the stimulation of voting. *American Political Science Review*, *20*, 869–874.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*, 765–789.

Jha, S., & Wilkinson, S. (2012). Does combat experience foster organizational skill? Evidence from ethnic cleansing during the Partition of South Asia. *American Political Science Review*, *106*, 883–907.

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Krauss, C. (2015). More oil companies could join Exxon Mobil as focus of climate investigations. *New York Times*, 6 November.

Lieberman, A., & Rust, S. (2015). Big oil companies united to fight regulations – but spent millions bracing for climate change. *Los Angeles Times*, 31 December.

Lincoln, A. (1897). *Political debates between Lincoln and Douglas*. Cleveland, OH: Burrows Bros. Co., 1897; Bartleby.com, 2001. Retrieved from www.bartleby.com/251/.

Lyall, J., Blair, G., & Imai, K. (2013). Explaining support for combatants during wartime: A survey experiment in Afghanistan. *American Political Science Review*, *107*, 679–705.

McDermott, R. (2002). Experimental methods in political science. *Annual Review of Political Science*, *5*, 31–61.

Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society*. New York, NY: Basic Books.

Paluck, E. L., & Green, D. P. (2009). Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda. *American Political Science Review*, *103*, 622–644.

Rogowski, R. (2013). Shooting (or ignoring) the messenger. *Political Studies Review*, *11*, 216–221.

Sekhon, J. S., & Titiunik, R. (2012). When natural experiments are neither natural nor experiments. *American Political Science Review*, *106*, 35–57.

Shami, M. (2012). Collective action, clientelism, and connectivity. *American Political Science Review*, *106*, 588–606.

Tomz, M., & Weeks, J. (2013). Public opinion and the democratic peace. *American Political Science Review*, *107*, 849–865.

RONALD ROGOWSKI SHORT BIOGRAPHY

Ronald Rogowski is a professor and former chair of political science at UCLA and served from 2007 to 2012 as lead editor of the *American Political Science Review*. He currently serves also part-time as visiting global distinguished professor at New York University, Abu Dhabi. He studied Political Science, Mathematics, and Sociology at the University of Nebraska, Princeton, Berlin, and Bonn. He has taught at Princeton, Duke, and Minnesota

and has held research appointments at Harvard, the Center for Advanced Study in the Behavioral Sciences, and the Wissenschaftskolleg zu Berlin. His principal books are Rational Legitimacy, Commerce and Coalitions, and (with Eric Chang, Mark Kayser, and Drew Linzer) Electoral Systems and the Balance of Consumer-Producer Power. He has worked most recently on (i) economic consequences of electoral systems (in addition to the book, articles in AJPS, JOP, BJPS); (ii) inequality and institutions, including a recent article on the economics and politics of slavery; and (iii) how the economic downturn of 2008–2009 affected the economies and the voting behavior of regions within the European Union. He has served as a vice president of the American Political Science Association and was elected in 1994 as a Fellow of the American Academy of Arts and Sciences.

RELATED ESSAYS

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure *(Methods)*, Robert Boruch and Alan Ruby
Repeated Cross-Sections in Survey Data *(Methods)*, Henry E. Brady and Richard Johnston
Network Research Experiments *(Methods)*, Allen L. Linton and Betsy Sinclair
Content Analysis *(Methods)*, Steven E. Stemler
Quasi-Experiments *(Methods)*, Charles S. Reichardt
Virtual Worlds as Laboratories *(Methods)*, Travis L. Ross *et al*.