The Experimental Political Scientist

In this issue

From the Editor

Welcome to the next issue of the Experimental Political Scientist. We have two new, hopefully recurring features: notes on across field collaboration and a book review of a recent book directly engaging with experimental political science. We also have two feature sections, one on bio-politics and one on the role of blocking and randomization tests. Finally, we have a follow-up on experimental protocol. Happy experimentation! Dustin Tingley, Harvard Government Department

Information on Joining or Contributing

The *Experimental Political Scientist* is the official newsletter of APSA Organized Experiments section 42. To receive the newsletters register for the section ($8/yr!) at http://www.apsanet.org and visit us at http://ps-experiments.ucr.edu/group. Graduate students and professors are highly encouraged to submit to the newsletter. This and the previous issues provide examples for submission themes. Questions can be sent to the editor. Deadline for the fall newsletter will be 10/15/2011.

Letter from the President: A Golden Era for the Gold Standard?

The experimental method is often referred to as the gold standard for causal inference. The founding of this section reflects disciplinary recognition that the experimental method has become a common part of political scientists' toolkit. Other signs of the penchant towards experimentation can be found in the dramatic rise, over the past decade, of published articles using the experimental method, and the publication of several books on experiments in political science. This should not, however, be taken as indicative that a golden era of experimental political science has arrived. Designing, implementing, and analyzing a good experiment brings with it a host of challenges that often escape notice, even to experimenters themselves. It is exactly now - as experimental approaches continue to spread through the discipline - that high standards must be upheld and educational efforts be undertaken. I am confident that this section and its contributors will take a leading role in such initiatives. Along these lines, the section has taken a number of steps to facilitate experimental research, keep members connected with the latest developments in experimental political science, and offer particular opportunities for young scholars. Here is a sampling of our activities:

A junior scholars committee that plans to institute a mentor match program at APSA so that graduate students and recent Ph.D.s can meet with more senior scholars to discuss their research; an experimental standards committee that is working to develop standards for reporting social science experimental research (the committee's suggestions will be distributed and discussed among section members); and a journal committee that is assessing the pros and cons for starting a new experimental political science journal. These efforts are reinforced by our newsletter and section website.

Of course the section also undertakes what all other sections do: organize APSA panels. This coming APSA will be the first in which the section has panels, thanks to the able organizational efforts of section head Kevin Arceneaux. Kevin received many more proposals than he could accommodate given the limited number of spaces allocated by the APSA. The best way for the section to receive more panels in 2012 is to come to the 2011 panels - APSA allocates slots based, in part, on attendance of the prior year's section's panels.

The section will not host a reception at APSA; instead, our plan is to use funds, as they grow, to set up small research grants for junior scholars' experimental research. Most such funds come from section membership fees, and thus, if you have not already joined, I implore you to do so. It only costs $8! I also encourage everyone to attend the section's business meeting at APSA where we will discuss work by the aforementioned committees and present book, paper, and dissertation awards.

This is an exciting time for experimental political science but one that requires our action and involvement. For the experimental method to play an increasing role in producing knowledge about government and politics, it is critical that steps be taken to ensure high standards.

The Odd Couple? Co-authoring across disciplinary boundaries.

Catherine C. Eckel
Economics, UT-Dallas
eckelc@utdallas.edu

Rick K. Wilson
Political Science, Rice
rkw@rice.edu

The Editor of this Newsletter asked us why we work together, and how we navigate the cross-disciplinary divide. We come from different disciplines and we speak (somewhat) different languages, but in our own work we use the same methodology. As a result, our skill sets are, to a large extent, substitutes rather than complements. Yet we have written a great deal together and have developed a common interest in questions that cut across numerous disciplines. Perhaps our experience has some relevance for understanding co-authorship more generally.

We first met at the Public Choice Society meetings in 1986. The Public Choice meetings once had a significant representation of experimental research, though that is less true now. Many economists interested in experiments attended these sessions. Because of the focus on collective decision making and voting, there also was a small contingent of political scientists, sociologists and the occasional psychologist attending. Both of us were presenting papers on strategic voting and until that point we had not realized we were working on a similar topic. Rick looked a little different in those days: he had very long hair and maybe even a beard - he's since abandoned the hippy professor look. Catherine's paper, with Charlie Holt, was published in the AER and cited Rick's paper. Rick's paper, with Bobbi Herzberg, was published in the JOP and Rick cited Charlie properly, but misspelled Catherine's name (you can look it up). After that we proceeded to ignore each other for 10 years.

We ran into one another at the National Science Foundation in 1996 where we were both starting as Program Officers for our respective disciplines. We were fortunate in that there was a large group of fellow Program Officers in our division who were also interested in using experiments. This prompted daily lunches and discussions of our own work among all of us. Like many co-authorships, this one started with one of us (Wilson) walking into the other's office (Eckel) and tossing a paper on the desk, pronouncing that the authors had done it all wrong. As usual, Catherine asked Rick to be specific, rather than general, and we quickly designed an experiment to test a key feature of the paper. Little did we realize that this was the start of a long-term co-authorship.

What Do We Each Bring to the Table?

With most co-authorships the co-authors each bring unique skills to the table. In this case it is not immediately obvious that this is the case. Both of us have used experimental designs for many years. Both of us have a passing acquaintance with theory and econometrics. As noted above, we appear to be substitutes rather than complements. Is this the best mix for scholars?

Despite the similarities, we speak different languages, we read different things, we approach questions differently, and we write differently. Catherine is an economist and Rick is a political scientist. While the language barrier is not huge, we have realized that in a co-authorship it is important to have an overlapping vocabulary sufficient to communicate effectively and precisely. Catherine often complains that Rick is not being sufficiently precise with a concept (and often she is correct). This leads to better organized papers and clearer implications for our findings. Our disagreement over language usually forces precision.

While we read a good deal of the same work, we are constantly egging one another on to read new articles. A typical email exchange might go: "You've got to take a look at this!" "Why? I'm busy." "You'll see." Sure enough, the article usually fits in with some old problem or serves as fodder for some new idea that we might want to tackle. Being able to share in our reading habits (running across neuroscience, biology, psychology, anthropology, and our own disciplines) is extremely fruitful. It allows us to quickly distill ideas and press forward on new topics.

We approach questions differently. Catherine, in the tradition of Economics, sees a problem, quickly figures out how it fits with the current disputes. Rick sees the problem and wants to immediately put it into the larger context. Catherine pinpoints the precise literature to which the question is addressed. Rick writes an intellectual history leading up to the problem, which Catherine then deletes. These are disciplinary differences and we have learned to navigate our stylistic differences.

We write differently. Rick's writing is bloated (this mostly is Rick's writing). It wrestles over various ideas and

approaches and often takes a while to come to a conclusion. Catherine's writing is focused and to the point. Over the years we have figured out how to best work together - Rick writes and Catherine cuts. The product is always better for the joint production.

So far we have detailed the differences. Where we have a common view is with experimental design. We take design seriously and often draw on each other even outside our own co-authorships. Four eyes are better than two when designing an experiment. We have avoided some major errors over time by thoroughly working over our designs.

Co-Authorships

One of the major concerns, especially for junior faculty, is the merit of co-authorships. Rick often hears senior faculty arguing that junior faculty must demonstrate their competence by producing single-authored work. Yet this seems to fly in the face of what is happening in the major journals. It seems that fewer single authored articles succeed. Multiple authored articles are becoming the norm in political science. In experimental economics (and economics more generally) multiple authorships are common. It is difficult to imagine how a single author can elaborate a theoretical contribution, design an experiment and carefully navigate the minefield of econometrics. Moreover, to successfully craft a paper that makes a contribution in more than one discipline is doubly difficult with a single author.

Here are some simple suggestions. Do not follow our example. It is useful to seek out co-authorships in which your skills complement those of others. Experimental design, as we all know, is exceedingly hard work. Getting the experiment right is imperative (without internal validity, we have nothing to say). Make certain that everyone knows the role you played (and it is probably with the experimental design). Of course you need to worry about the theoretical underpinnings of the experiment. And you need to worry about the empirical component of the paper. But, your contribution should be clear.

Do not be afraid to step across disciplinary boundaries. Well, be a little afraid. It is important to be clear about what you expect to get from such a relationship, especially if you are untenured. Will your Department value consorting with an Economist (or, gasp, a Political Scientist)? Will your Department understand the journal that you publish your experiment in and how will it count that contribution? In many economics departments, publications in other fields - even in Science or Nature - just won't count. Economists are narrow minded that way. Will your Department count citations outside the discipline? These are questions that you should ask. Eventually disciplines will learn to value cross-disciplinary work. However, there is some burden on you to make it clear the contribution to your own discipline.

While it can be very rewarding to work with a senior scholar in your or another field, we advise that you seek out co-authors who are peers. It is also good not to work all the time with one coauthor, as that makes your contribution harder to distinguish come tenure time. Rather than working with a senior faculty member, develop a cohort of people who share common interests. In this manner you won't be seen as simply the experimental "hired gun" who has no ideas and makes no contribution outside the lab. (We should note that the hired gun approach has been very successful for one or two people we know, so nothing is absolute.) It is important to signal ownership of an idea, best achieved by having evidence of a research program, with or without coauthors. But coauthors, especially those outside your own discipline, can help develop your research agenda and highlight its relevance.

Long-term co-authorships can be very rewarding. Catherine and Rick have worked together for some time. Yet both of us have other co-authors we have worked with for even longer periods of time. Long-term co-authorships can come to resemble old marriages - we often finish one another's sentences. An added benefit is that, at conferences that we jointly attend, we always have a dinner partner to depend on.

What We've Learned From One Another

Rick has gotten a great deal out of this collaboration. Prior to working together Rick had a rigid view of theory, hypothesis construction and the role of experiments in tackling questions. Catherine got Rick to question anomalous findings and to be willing to go beyond the narrow confines of political science. Catherine was Rick's introduction to Behavioral Economics.

Catherine is surprised to hear this. When we were at NSF we funded (jointly or in collaboration with other programs) the first neuro-economics experiments, the first internet experiments, the first economics experiments with kids, and lots of stuff that crossed disciplinary boundaries. We were a bit notorious for championing "risky" research

with mixed reviews. Together we read Tooby and Cosmides, Daly and Wilson, Pinker, Matt Ridley, and others, and for the first time began really to think about connections between biology and economic/political behavior. This sure was a lot of fun! And Catherine is pretty sure she wasn't the leader in all this!

We develop our ideas together, and we design together. Rick generally supervises the programming, or does it himself. Rick is also absolutely great at the first draft! For Catherine, writing is slow and deliberate, but after a long conversation, Rick can whip out a first draft in minutes!

In the conferences we attend, and in the summer courses we teach for PhD students in political science, Rick likes to cultivate a persona that is grumpy and critical. He fools a lot of people. But underneath all that he is much more patient and kind than Catherine. In working - and especially in teaching - with him, Catherine has learned to listen more carefully and patiently, and to extract experimental designs from the fuzziest of student ideas. Still it is true that Rick spends more time and is more generous with his groups that she is with hers.

At the moment Catherine wishes that Rick were not editing the AJPS. He says it "only" takes him four hours a day. We haven't written a new paper in, what, three years? Oh dear, time to get back to work!!

Future of Experimental Approaches to Biopolitics

Rose McDermott
Brown University
Rose_McDermott@brown.edu

Experimental tests of political phenomena have achieved increasing relevance and credibility over the last decade, largely lead by developments in American politics related to the investigation of voting behavior. More recent extensions of field experiments into this arena both in the United States and elsewhere have continued to expand this trend. Meanwhile, extensive experimental tests of the influence of biology, physiology and genetics on complex social behavior have taken place outside the field, but few of these applications have related to politics. Attempts to join these two traditions, while rare, represent an important opportunity for those wishing to undertake experimental investigations into new venues in political science.

The standard dependent variables that tend to be explored by political scientists, specifically those related to voting behavior, remain unnecessarily restrictive. Although such variables constitute the bread and butter of the field, and certainly represent important and easy variables to investigate experimentally, other large, complex and important social and political variables can, and should, be investigated experimentally as well. The most obvious exist at the overlap of behavior economics and comparative or international political economy, including topics quite amenable to experimental exploration, including foreign direct investment and foreign aid, perceptions of monetary and welfare policies, and issues surrounding trade imbalances and dependencies between allies and adversaries. Although many of these issues may not appear to be obviously investigated from the perspective of biological and genetic issues, they are in fact conjoined at an underlying conceptual level by theoretical suppositions regarding the nature of in-group and out-group relations, as well as larger topics related to the nature of cooperation, trust and aggression.

Political science as a discipline fundamentally revolves around the examination and exploration of issues of power, just as psychology concentrates on those related to affiliation and economics focuses on those involved in money. As a discipline, we might be best served by playing to our strengths, and working to develop experimental tests of those substantive areas related to our disciplinary strength, while working in collaboration with those in other fields, to explore the basis of dominance and power from a biological and genetic substrate. It would not be surprising, given what recent studies have shown in the primate literature, to discover the phylogentic roots of dominance in ancient instantiated psychological mechanisms. Recent work has been able to demonstrate these phenomena in macque monkeys using an adapted version of implicit association tests. Such work might also illuminate the ways in which particular local geographies and ecologies may potentiate individual and regional population variances across genotypes, as the work of Joan Chaio has shown in her work on status and dominance across regions. Using genetic assays and functional magnetic resonance imaging, increasing number of people in social neuroscience and other areas have begun to explore these topics in very innovative ways, using sophisticated experimental paradigms.

Experimental examination of topics related to voting

and economics, just like those related to dominance, while crucial for our understanding of important political behavior, rely in critical ways in theories, methods, and models derived from other fields. This sort of collaboration, while not new, needs to transform from the traditional borrowing model favored by political scientists who tend to use theories and methods developed in other fields to illuminate substantive issues of concern in political science, to more actively contribute our knowledge to broader interdisciplinary and collaborative teams exploring issues of substantive interest together. Political science can bring a strong method to the table by incorporating the use of surveys and extended interview protocols in the experimental exploration of the biological basis of complex social and political behavior. These tools can be used to help contextualize the relationship between genotype and phenotype within critical developmental pathways. This strategy can help illuminate the variety of crucial environmental triggers which can contingently cue particular behavioral repertoires, many of which have been extensively explored in previous work in political science. This conjoining of past substantive theoretical work and methodological strategies in political science with novel explorations of the biological contributions to instigating, maintaining and curtailing complex social and political choice opens up an important new avenue for experimental exploration. Such a research agenda can help enlighten the complex ways in which biology, environment and culture reciprocally interact to encourage some behavior while extinguishing others, and provide insight into how and why individuals differ in their ability and willingness to direct and control others to join together to cooperative achieve goals which range from construction to destruction.

Chris Dawes
University of California at San Diego
cdawes@ucsd.edu

The genopolitics literature took a major step forward when Alford, Funk, and Hibbing (2005), motivated by earlier work published by prominent behavior geneticists, demonstrated that variation in political ideology could be attributed to genetic variation. Scholars have since found evidence of heritable variation in a wide variety of political behaviors and attitudes including voter turnout, strength of partisan attachment, political efficacy, political knowledge, interest in politics, and measures of overall political participation. The work-horse research design employed by these studies relies on comparing traits of sibling pairs with differing genetic relatedness to back out estimates of heritability. Consequently, they do not tell us which genetic variants are correlated with political behaviors and attitudes but rather the influence of genes in total.

A subsequent wave of genopolitics research has sought to identify specific genetic variants that are correlated with political behaviors. This can be done using one of two approaches. The first is a theoretical approach that focuses on ”candidate” genes based on hypotheses about how the known functions of the genes could be related to the mechanisms underlying political behavior. Thus far, political scientists have focused on genes previously shown to be related to traits such as pro-social attitudes and behavior. The second approach, known as a genome-wide association study, interrogates hundreds of thousands or millions of variants located across the genome looking for a correlation between each of these variants and a given political trait.

I would argue that the next wave of genopolitics will be focused on extending this earlier work in two related areas. First, while knowledge that particular genes are correlated with political traits is interesting, the causal pathways linking the two are ultimately of more theoretical value. Recent work suggests that personality traits, other-regarding preferences, temporal discounting, attitudes towards risk, and cognitive ability may mediate the relationship between genes and political traits. However, most of these potential mediators have yet to be formally tested and researchers have had to rely on observational data in the few tests that have been conducted. Yet, as Bullock, Green, and Ha (2010) point out, mediation analysis using observational data is likely to be biased (though see Imai et. al (working) for discussion of new experimental design based approaches and sensitivity analyses that may be of some help). While experimental methods are not a panacea, they offer the best opportunity to illuminate potential causal pathways.

The second expansion of the genopolitics literature will be an explicit consideration of the ways in which genes and the environment work together to influence political traits. A limitation of the standard twin or association study designs is that they tend to rely on an overly simplistic functional form that assumes the influence of genes and environment are additive and independent of one another. However, it is more likely that genes and environmental factors combine

interactively; our genetic endowment likely influences the types of environmental settings we are exposed to either because they influence which ones we seek out, or which are imposed on us. The former scenario is known as gene-environment interaction and the latter is known as gene-environment correlation. A handful of published political science studies have explored gene-environment interactions; however, since they rely on observational data, it is possible that exposure to the environmental factor being studied is not truly exogenous but is due, at least in part, to gene-environment correlation.

Field experiments offer an excellent way to disentangle these two phenomena. For example, Jaime Settle, Peter Loewen, and I found based on observational data that individuals with a certain version of a gene previously found to be associated with sensitivity to stressful stimuli were less likely to vote when living in a more competitive political environment, as compared to those with the alternative variant. Our hypothesis was that a genetic predisposition to stress sensitivity combined with a politically stressful environment caused citizens to disengage from politics. However, a problem with our research design was that individuals sensitive to stress may choose to avoid living in such environments. In order to overcome this problem, we conducted a field experiment with Costas Panagopoulos in the 2010 California election that randomly assigned a mobilization message designed to elicit the type of stress we hypothesized in the previous study helped cause disengagement. The fact that the treatment was randomly assigned, and could be compared to a control group, means that the environmental influence was truly exogenous. While this type of experiment has previously been done with measures of personality traits, our study is an example of how experimental methods can be used in the study of genopolitics.

Political scientists have increasingly been granted access to large genetically informative samples and are also actively creating their own novel samples, making field experiments even more feasible in the future. Recent work has also conducted laboratory experiments on genetically informative samples. In the foreseeable future, experiments will be the norm in better understanding the causal relationship between genes, the environment and political traits.

Alford, J., C. Funk and J. Hibbing. 2005. Are Political Orientations Genetically Transmitted? American Political Science Review 99(2):153167.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. Yes, But Whats the Mechanism? (Dont Expect an Easy Answer). Journal of Personality and Social Psychology 98 (April): 550-58.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. "Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies." Working paper available here

The Role of Protocol in the Design and Reporting of Experiments

Thomas J. Leeper
Northwestern University
leeper@u.northwestern.edu

An average treatment effect is relatively simple to estimate. The basic requirements of experimental research that make that estimation straightforward and unconfounded – manipulation of causal variables and control over all else – are often far more difficult to achieve than is often imagined or implied. In this article, I continue the discussion started by Lupia and Druckman in the inaugural issue of this newsletter by emphasizing how the maintenance and publication of a detailed experimental protocol is critical for the analysis and reporting of experiments.

Developing Experimental Protocol

Like any novel data collection, an experiment is complicated. It begins with research questions, theory-building, and hypothesis generation. Experiments are unique, however, in that these initial stages are followed by a somewhat murky process of developing stimulus (treatment) materials, disseminating those stimuli, and assessing outcome measures. The idea of "just doing an experiment" has probably occurred to many political scientists – including non-experimentalists – but the process of executing an experiment, and doing it well, is rarely a topic heavily emphasized in published descriptions of experiments. Executing an experiment well requires quality protocol, by which I mean the detailed justification and explanation of the experiment that will be implemented and the record of how that intended experiment may not have occurred, for what units, and why.

A well-thought-out experimental protocol is the experimenter's version of the observationalist's well-thought-out statistical model of causes, outcomes, and controls. Rosenbaum (2010) describes a "poorer observational study"where "if sufficiently many analyses are performed, something publishable will turn up sooner or later." This is not the way research – experimental or observational – should be conducted. Instead, he offers that "before beginning the actual experiment, a written protocol describes the design, exclusion criteria, primary and secondary outcomes, and proposed analyses" (7). Lupia (2010) suggests that researchers record "Record all steps that convert human energy and dollars into datapoints." Both quotes provide useful general advice, but the experimental protocol should also includes many specific features about the intended experiment, including:

Theory and Hypotheses

- Details of what outcome variable is under examination (including the values of its potential outcomes) and the causal factors (alone or in combination) that are theorized to affect the outcome variable
- References to relevant prior designs (published or unpublished) that inform the design, implementation, and/or analysis of this experiment
- Exact listing of hypotheses with discussion of how each hypothesis is reflected in one or more features (manipulations or controls) of the intended design
- Discussion of how hypotheses were used to construct the design (rather than design to construct hypotheses), including how each hypothesis is testable (i.e., falsifiable) in the intended design, including any anticipated confounding between manipulated causes and observed or unobservable alternative causes

Instrumentation

- Details of how theoretical constructs are manipulated by the experimenter, including exact wording and presentation of stimulus/treatment materials
- Details of how covariates and outcome variables are measured and scaled and exact question wordings or coding schemes if subjective and/or self-report measures are used
- Explanation of how stimuli are predicted to affect only the intended causal construct[1]
- Details of pretesting (of stimuli and/or outcome measurement techniques) to validate whether stimuli manipulate the causal constructs as anticipated by the researcher, by manipulating only the intended cause, in the theorized direction, and with what intensity
- Intended mode(s) in which each unit will be exposed to each stimulus (e.g., in-person, via phone, internet, mail, television, radio, print, etc.) and the mode in which outcomes will be observed (e.g., in-person, laboratory, internet, phone, voting records, etc.)

Population, Sample, and Treatment Assignment

- Details of sample construction, including how a sampling frame was constructed of population units (if applicable) and how units from that frame were selected for or excluded from treatment assignment
- Details of randomization procedures, including how random numbers were generated (whether a uniform or some other probability distribution was used and from where those numbers were obtained) and assigned (across the entire sample, or within blocks, clusters, or a combination thereof)

Implementation

- Intended (and executed) schedule for when, where, and how treatments will be applied and outcome variables measured and by whom (the experimenter or an assistant, or a government agency, corporation, etc.); if multiple sessions or repeated exposure to stimuli are part of the design, the schedule should specify and justify the order of what stimuli are applied at each point in time
- Procedures for how units are made blind to their treatment assignment, how data are collected and stored regarding each unit prior to analysis, and how

[1] If experimenters are uncertain about how a given stimulus will alter the theoretical construct, pretesting should be conducted (as described). If a stimulus potentially affects more than one causal construct (and alternative stimuli are infeasible), nonequivalent outcome measures – that measure the desired construct and separately measure other causes that should be unaffected by stimuli – should be developed to demonstrate that the intended cause produced the observed outcome and a confounding cause was not at work.

anyone implementing the study are blind to each unit's treatment assignment and value(s) of outcome variable(s) until analysis

- Details of manipulation checks and post-randomization covariate balance tests, including how and when manipulation checks will be administered and what balance tests will be used and what predetermined degree of imbalance will be considered problematic for experimental inference
- Procedures for how deviations from protocol will be recorded and utilized in data analysis, including errors made by experimenters, item and unit noncompliance or attrition, and other relevant notes

Analysis

- Definition of specific causal effects to be estimated for testing each hypothesis; if more than two treatment groups are implemented, this should include what treatment groups or pooled treatment groups are used to estimate those effects (i.e., specification of contrasts)
- Explanation of specific statistical analysis to be performed, including how covariates will be used in the analysis if at all (i.e., for regression, subclassification, subgroup analysis, etc.)
- Plan for how to analytically address noncompliance, attrition, missing data, or covariate imbalance (through dropping observations, multiple imputation, intent-to-treat analysis, etc.)
- Additional analyses or changes to this protocol should be recorded in a "lab book"(Lupia 2010) and a record should be kept of all analysis performed (in order to reduce the risk of Type I errors)

In the same way that a regression model includes putatively causal variables and "controls for" covariates and confounds, the experimental protocol establishes what factors vary between experimental units and what factors are "controlled for" by uniform execution of the experiment across randomized units. The protocol determines what features (of subjects, of context, and of experimenters) are included (observed and analyzed), excluded (unobserved), or controlled for (observed or held constant across units) in the final analysis, even if that analysis only consists of a simple difference of means or difference of proportions test for the average causal effect. Thinking about protocol forces the experimenter to make and record decisions about what is controlled in the experiment decisions that might otherwise not be made, leading to unclear control or a loss of control due to uneven implementation of the intended design.

Deviations from Protocol

The protocol should include description and justification of all decisions regarding the *intended experiment* and *implemented experiment*: i.e., if cost, feasibility, sample size, time, ethics, or other constraints alter the intended procedure, a record of these changes in the protocol document (and the justification thereof) is the only way to fully communicate what choices were made by experimenters and why. Druckman (2010) noted that experimental research is characterized by several myths; the myth of experiments always working perfectly is probably a necessary addition to his list; all experiments are – at some level – broken experiments. The experimenter's concern needs to be placed on how to prevent, record, and compensate for the deviations from protocol that manifest in the implemented experiment. Compensating for deviations that have already occurred is more difficult than preventing them, but can only be done when deviations are anticipated and properly recorded.

Rather than pretend that deviations from protocol do not occur, researchers should keep the "lab book" called for by Lupia to record statistical procedures, but they need to also record the earlier unit-, treatment group-, and sample-level deviations from the intended experiment protocol before any statistical analysis is performed. These deviations (especially if they correlate with treatment group assignment) may undermine the elegance of experimental inference by confounding causal factors and altering other aspects of experimental control. At a minimum, if any variations in implementation emerge among units, unit-level variables should to be incorporated into the experimental dataset that record:

- the date, time, and location that stimuli were distributed or applied
- the date, time, and location that outcomes were measured
- location and/or mode of each stage of the experiment

- waves of implementation, which place units in different broader contexts even if procedures in each wave are identical
- exogenous contextual events that influenced some units but not others (due to timing of implementation, geographical locale, etc.)
- any changes in questionnaires, changes in stimuli, or changes in the contact between experimenter and experimental units that affect only some units
- different experimenters/interviewers/coders/assistants who administered stimuli, collected outcomes, or coded data[2]
- other research-relevant, unit-level or group-level deviations from protocol[3]

Reading a typical journal article, it would seem the types of deviations from protocol included in the above list are rarely seriously anticipated by researchers, considered in data analysis, or honestly reported in publication. Dropping observations and/or conducting intent-to-treat analysis are common answers to deviations from protocol (often for noncompliance, application of the incorrect treatment, or attrition), but these are only applicable in some cases and each may introduce bias into estimates of causal effects depending on the type and extent of deviations. Other answers certainly exist, but before we can address these problems, we need to hold ourselves to high and uniform standards of adopting, recording, and disseminating our experimental protocols (see Gerber, Doherty, and Dowling n.d.).

Reporting Protocol (and Deviations)

This level of full disclosure of experimental protocols is relatively rare.[4] While a journal article will often include a brief discussion of protocol, specific details of protocol (and deviations) are often omitted to save space or because those details are seemingly unimportant (especially in the eyes of non-experimenters). But, this is unfortunate and often reflects, at best, partial compliance with standards for reporting observational research. Druckman (2010) points out "these standards [for surveys] do not ask for the reporting of information critical to experimentation, such as randomization checks, manipulation checks, pre-test results" (10). Reporting on experiments requires more thorough description in order for consumers of that research to understand how the causal effect(s) might be constrained by particular context, unit characteristics, poorly implemented protocol, and so forth.[5]

Including the full protocol as an (online) appendix is better than providing insufficient detail in an abbreviated methods section. Reviewers would criticize a paper with multiple statistical models if only one model were fully described and would certainly take issue with analysis that did not clearly specify the variables included in the analysis and the rationale for controlling for or excluding particular factors. Presenting a detailed experimental protocol is the experimenter's means of justifying relatively simple statistical analyses. But an average treatment effect is not just a difference in expected outcomes between groups; it is a difference conditional on a large number of features of the experimental design and broader context that are only clear if they are considered, recorded, and explained by the experimenter in the protocol document.

Seemingly similar experiments that yield different effects estimates or substantive conclusions – think cold fusion! – may be due to dissimilarities in protocol and implementation that fail to be noted in lab books or reported in published results. Experimental researchers need to commit to incorporating deviations from protocol into their final datasets, to sharing of those data, and to publication of full protocols for the purposes of replication, meta-analysis, and scholarly critique. Reproductions of stimuli, exact questionnaires (or other outcome measurement tools), raw datasets, statistical syntax files, and detailed protocols are all necessary addenda to any published experiment.

Conclusion

[2]Demographic variables for these individuals may need to be included if they might influence unit behavior on sensitive issues (such as race or gender for studies that examine those issues) or if data were gathered on each unit in systematically different (but unobserved or unobservable) ways.

[3]This could include units expressing anger or anxiety about the researcher's intent in a study of emotions in politics, units communicating with other units (which violates SUTVA), units that receive a treatment other than the one assigned or were exposed to more than one treatment, etc.

[4]Although social scientists rarely share their protocols, online sharing of protocols in the biological sciences is common at sites such as Nature Protocols (http://www.nature.com/protocolexchange/) or SpringerProtocols (http://www.springerprotocols.com/).

[5]Given that experimental research enables meta-analytic review, thinking as a meta-analyst can also help researchers determine what protocol information to report for a given study (see Lipsey 1994; Stock 1994).

The broad agenda of any experimental science is replication and complication – the progressive addition of moderating factors, search for mechanisms, and exploration of alternative causes and outcomes. Writing and reporting protocol is a critical best practice to establish strict experimental control, to minimize noncompliance and uneven implementation, to produce clear and accurate presentation of results, and to properly interpret and generalize experimental findings. Given the importance of the (often unwritten) information contained within (often unpublished) protocol to progressive scientific inquiry (including replication and critique), we should not be afraid of producing and sharing detailed records of our scientific endeavors. But, more importantly, we should not forget the importance of producing and enforcing protocol for preserving the elegance of experimental inference.

References

Druckman, James N. 2010. "Experimental Myths." *The Experimental Political Scientist* 1(1): 9-11.

Gerber, Alan S., David Doherty, and Conor Dowling. 2009. "Developing a Checklist for Reporting the Design and Results of Social Science Experiments." Presented at Experiments in Governance and Politics Network meeting, Yale University, April 24-25, 2009.

Lipsey, Mark W. 1994. "Identifying Potentially Interesting Variables and Analysis Opportunities." In Harris Cooper and Larry V. Hedges, eds., *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Lupia, Arthur. 2010. "Procedural Transparency, Experiments and the Credibility of Political Science." *The Experimental Political Scientist* 1(1): 5-9.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.

Stock, William A. 1994. "Systematic Coding for Research Synthesis." In Harris Cooper and Larry V. Hedges, eds., *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Review of From Nature to the Lab: Experimental Political Science and the Study of Causality by Rebecca B. Morton and Kenneth C. Williams

Sean Gailmard
Charles and Louise Travers Department of Political Science
University of California at Berkeley
gailmard@berkeley.edu

The experimental community in political science has reached a level of self-awareness that seemed like a pipe dream only a few years ago. Correspondingly, experimentalists have confronted existential questions such as, Where did we come from? How did we get here? What is our purpose? And what are our sacred beliefs?

From Nature to the Lab covers all this ground in its review of the study of causation, experimental methods, and their intersection in empirical political science research. It is an ambitious, and in my view, successful book. Morton and Williams cover the theoretical link between experimentation and causal inference; the issues of validity that scholars face when using experimental and observational data; a wealth of practical advice about the business end of conducting an experiment (including a narrative "checklist" in the appendix); and a wide range of research questions addressed, experimental designs used, and results obtained by experimentalists in political science over the years. The book is conceptually and substantively rich but not technically demanding. It should be accessible to first year graduate students with one prior course on statistical methods and inference; the text would easily occupy a one-semester course on experimental methods. The book will also be useful to practicing researchers new to experimental methods, and enriching to seasoned experimentalists.

Moreover, despite their pedigree in one of political science's several experimental traditions (namely, the game theoretic/economic tradition), Morton and Williams are self-consciously ecumenical in their coverage of issues and literature from multiple traditions (namely, psychological as well). This is not a book aimed at one "tribe" of experimentalists, or at converting members from the other tribe who have lost the true way. Rather, the book covers theoretical and practical issues of common interest to all

experimentalists, as well as frankly reflecting a few points of debate ($e.g.$, subject incentives, deception).

The book is divided into three substantive sections. The first covers the theory of causation in terms of the Rubin causal model, the nature of experiments (which are defined in terms of purposive manipulation of treatment variables), prospects for control of confounding variables in experimental and observational research designs, and what the authors call the "formal theory approach" to causation.

One of the most important aspects of this manuscript is that it thoroughly ties experimentation to problems of causal inference in a comprehensive and systematic way. Given the convergence of many social science disciplines on attention to rigorous causal identification, and the folk wisdom that experimentation is the gold standard of causal inference, this link may seem natural or obvious. Yet this focus is either lacking or implicit in many (especially early) political science and economics discussions of experiments. Many early experiments, especially in the game theoretic tradition, can appear more as a sort of survey—what do real actually do in a particular situation, and do they play subgame perfect equilibrium strategies?—than an attempt to identify causal effects of treatment variables. From a Rubin model standpoint, for instance, it is not clear what causal effect is established by the first ultimatum game experiments, or experiments on any particular game in isolation.

A central premise of the book's presentation on causal inference is that there is no causal identification without theory. In the sense that any model of causation, including the Rubin causal model, makes theoretically contestable claims about the meaning of causation, this is surely true. But it does not follow that, having accepted the Rubin causal model, additional substantive theorizing about the relationship between treatment and outcome is necessary to identify a causal effect in an experiment. Just as the treatment effect of aspirin on headache pain was known long before any coherent theory about the reason for this effect was available, an experiment can identify the effect of negative campaign advertising on interest in an election without any particular psychological theory explaining the effect. To be sure, such a theory would be useful to determine which conceivable experiments would be worth conducting, but that is a separate question from the one of interpreting the effects of any given treatment in whatever experimental data happens to be available.

Though *From Nature to the Lab* is explicit about its connection of experimentation and causality, it does not entirely accept the discipline's convergence on the Rubin causal model (RCM) as a touchstone of causation. This is most apparent in chapter 6, "Formal Theory and Causality." Here Morton and Williams flesh out the "formal theory approach" (FTA) to causality. They contend that the key difference between the RCM and FTA approach to causal inference in experiments is this: in the RCM approach, a researcher assumes without proof that assumptions underlying the causal prediction evaluated are consistent with the assumptions underlying the method used to evaluate them (p. 198); in FTA, a researcher "carefully investigates whether each assumption for the theoretical model holds for the empirical analysis and how these assumptions matter in that analysis" (p. 202). This would include, for instance, the assumption that the payoffs instantiated in an experimental game match the utilities of the relevant abstract representation ($e.g.$ the extensive form of that game).[6]

This is a high bar, and indeed it is not clear it is attainable. How can a researcher ever be completely sure that the subjects' perception of a game coincides exactly with the formal representation in the underlying theory? And if we are not completely sure, how can the concordance of actual behavior in a game as instantiated in an experiment and equilibrium "predictions" be taken, *in and of itself* and without comparison to any other game, to supply causal information? Such a concordance seems inherently vulnerable to arguments of confounding, from an RCM standpoint. For instance, if an exceedingly complex game is so inscrutable to subjects that they choose randomly (or choose the first option offered, etc.), and this happens to coincide with the equilibrium prediction in that game, what can be inferred about the causal effect of the strategic structure on actual behavior?

This very question, of course, assumes that the RCM standard of causation can be used to evaluate causal claims in the FTA framework. Morton and Williams do not attempt to "square the circle" of reconciling the causal pretensions of each approach within a single coherent theory of

[6]In Morton and Williams's framework, experiments on formal theories may be couched in either an RCM or FTA framework; non-formal theories are consigned to RCM frameworks only. This implicitly elevates the rhetorical powers of mathematics above mere verbal expressions of logic: the Morton-Williams FTA framework implies that expressing the exact same logic of a formal model in verbal terms without the symbolic representation loses something.

causation—an interesting question for methodological research, perhaps, and therefore an unreasonable goal for a book of this nature. Instead, and consistent with the tenor of the book more generally, they take each tradition on its own terms and explore the logic of causal identification in experiments within that tradition.

The second core section covers the theory of validity and its implications for experimental and observational research designs. Morton and Williams adhere faithfully to the updated validity typology of Shadish, Cook, and Campbell (2002). They provide a spirited argument against the conventional view that experiments, while achieving maximal internal validity, score lower on external validity than observational designs. They maintain, instead, that external validity is essentially an empirical question, and thus anyone arguing against the external validity of experimental designs must adduce evidence before this conjecture can be accepted. Moreover, Morton and Williams contend that one cannot logically analyze external validity of experiments in situations where internal validity (including statistical, causal, and construct validity) is not also satisfied. Together, these empirical and logical arguments present a nearly insurmountable burden for the hapless critic who assails the generalizability of experimental findings.

Despite this argumentation, one supposes that the critique of experiments in terms of a home-grown notion of "external validity" will continue unabated, and therefore experimentalists would do well to consider the verisimilitude of the environments they create. In the end, Morton and Williams are sanguine to this, and present analyses of artificiality of experimental environments (including subject pools and incentives) that experimentalists can draw on in design and presentation of results.

The final substantive section addresses the ethics of experimentation. This discussion covers both the history and current content of standard codes of ethical conduct in experiments, and also the issue of deception, one of the principal points of contention between various experimental traditions in political science (game theoretic vs. psychological). This is arguably the area where Morton and Williams get closest to advocacy of the norms of one experimental tradition over another. Yet their presentation of the costs and benefits of deception is quite even-handed, and their call for a norm of minimal deception eminently reasonable—particularly as pertains to institutions where researchers from the two dominant traditions share a subject pool.

Overall, *From Nature to the Lab* gives a comprehensive overview of the theoretical benefits of, and theoretical and practical issues involved in, conducting and analyzing data from an experiment. It is at once a theoretical analysis of, primer on conduct of, and review of literature from political science experiments. Though the book certainly contains the sorts of contestable claims that will make it interesting to experimentalists and political methodologists, its judicious treatment of an emerging canon of theoretical and practical issues in experimentation will make it a standard text on the topic for years.

Does Blocking Reduce Attrition Bias?

Thad Dunning
Department of Political Science, Yale University
thad.dunning@yale.edu

Several political scientists have recently drawn attention to the merits of blocking in experimental studies.

Experiments allow for unbiased estimation of average causal effects for the experimental study group, as long as attrition and some other threats to internal validity do not arise. However, estimators may be more or less precise. This is where blocking comes in. Here, units are grouped into strata, or blocks, and then randomized to treatment and control conditions within these blocks.

For instance, experimental units may be grouped into pairs with similar incomes, past voting histories, or values of other variables that may predict an outcome. In such "matched pair" designs, one member of each pair is randomly assigned to

treatment, while the other is randomly assigned to control.

As Moore (2010) and others have pointed out, blocking can be a worthwhile strategy, especially in smaller experiments. Blocking is most beneficial when units are relatively homogenous (with respect to the outcome) within blocks and heterogeneous across blocks. Thus, if investigators can identify variables that are good predictors of the outcome, blocking units before randomization may increase the precision of treatment effect estimators.

Matched-Pair Designs and Attrition Bias

Yet, can blocking overcome bias from attrition or non-compliance? Following King et al. (2007), Moore (2010) suggests that "Through blocking, design can anticipate and overcome a frequent field experiment reality: some units may be compromised during an experiment, and they and their blockmates can be excluded from analysis without endangering the entire randomization."

Attrition occurs when some units are lost to follow-up—that is, when the outcomes for some units are not recorded, for a variety of possible reasons. Such attrition can be devastating to causal inference when it is associated with treatment assignment and is a function of potential outcomes (that is, the outcomes that we would observe for each unit if it were assigned to treatment or control).

For example, the subjects most likely to drop out of a medical trial may be those assigned to treatment, for whom the treatment does not appear to be working. In this case, comparing health outcomes of subjects who remain in the treatment group to subjects assigned to control may overstate the efficacy of treatment, since subjects for whom the treatment did not seem to work dropped out of the trial.

Similar issues can arise with non-compliance, which occurs when units assigned to treatment receive the control, or vice versa. For instance, in randomized public policy programs, politicians may be tempted to administer the treatment to control units (King et al. 2007). If we conduct analysis by treatment received—comparing those who receive the treatment to those that receive the control—we risk bias. This is because units that crossover from control to treatment, or vice versa, may be unlike those that don't, in ways that matter for outcomes.[7]

Yet if attrition or non-compliance are functions of potential outcomes, dropping blocked pairs does not eliminate the bias. To see this, imagine that there are two types in the experimental population:

(1) Units that will be compromised (e.g., be subject to attrition) if they are assigned to treatment; and

(2) Units that will not be compromised, whether assigned to treatment or control.

If subjects are randomized in pairs, and we throw out all the pairs in which the treatment unit is compromised, then we know that all of the treatment units that are thrown out are of type (1).

Yet, the control units in the excluded pairs may include both types (1) and (2). For these control units, we don't get to observe the counterfactual response to treatment assignment—that is, whether the control unit would stay in the experiment or not if assigned to treatment.

This can lead to bias, if the propensity to be compromised is related to potential outcomes, because the control group contains types (1) and (2), while the treatment group contains only type (1).

A Numerical Example

A numerical example may help to illustrate this point. Suppose that in a small experiment, there are four units that will be blocked on a single covariate and then randomized to treatment within blocked pairs.

The table below shows the covariate value for each unit; the potential outcome under control, which is denoted $Y(0)$; and the potential outcome under treatment, which is denoted $Y(1)$. The final column records whether or not the unit will be lost to attrition if assigned to treatment.

The final row of the table shows the mean outcome if all units were assigned to control and the mean outcome if all units were assigned to treatment. The average causal effect is defined as the difference between these quantities.

[7] Non-compliance is typically not as harmful for inference as attrition: we can always do intention-to-treat analysis, which estimates the effect of treatment assignment. There are also sensible procedures for estimating the effect of treatment on compliers, using treatment assignment as an instrumental variable.

Covariate value	$Y(0)$	$Y(1)$	Lost to attrition if assigned to treatment?
2	1	2	NO
2	3	3	NO
3	4	6	YES
3	2	3	NO
Mean outcome:	$\frac{10}{4}$	$\frac{14}{4}$	

Here, the true average causal effect is $\frac{14}{4} - \frac{10}{4} = 1$.

Now, suppose we randomize in pairs defined by values on the covariate: for instance, we flip a coin to decide which of the two units with covariate value 2 goes into treatment, and similarly for the two units with covariate value 3.

This implies that there are 4 possible random assignments. The realized outcomes under each assignment are listed in the tables below. The pairs that will be dropped due to attrition are indicated with diagonal slashes.

For each randomization, the estimated average causal effect—that is, the mean outcome in the treatment group minus the mean outcome in the control group, after dropping the pairs in which there was attrition—is listed below each table.

	Covariate value	Realized outcome (control units)	Realized outcome (treatment units)
Randomization 1	2		2
	2	3	
	~~3~~		~~6~~
	~~3~~	~~2~~	
	Mean outcome:	$\frac{3}{1}$	$\frac{2}{1}$

Estimated average causal effect: $2 - 3 = -1$

	Covariate value	Realized outcome (control units)	Realized outcome (treatment units)
Randomization 2	2		2
	2	3	
	3	4	
	3		3
	Mean outcome:	$\frac{7}{2}$	$\frac{5}{2}$

Estimated average causal effect: $\frac{5}{2} - \frac{7}{2} = -1$

	Covariate value	Realized outcome (control units)	Realized outcome (treatment units)
Randomization 3	2	1	
	2		3
	3	4	
	3		3
	Mean outcome:	$\frac{5}{2}$	$\frac{6}{2}$

Estimated average causal effect: $\frac{6}{2} - \frac{5}{2} = \frac{1}{2}$

	Covariate value	Realized outcome (control units)	Realized outcome (treatment units)
Randomization 4	2	1	
	2		3
	~~3~~		~~6~~
	~~3~~	~~2~~	
	Mean outcome:	$\frac{1}{1}$	$\frac{3}{1}$

Estimated average causal effect: $3 - 1 = 2$

The average estimated causal effect is the average over all of the estimates obtained after each of the 4 possible randomizations, that is,

$$\frac{-1 - 1 + \frac{1}{2} + 2}{4} = \frac{1}{8}.$$

This estimator is biased by about -88%, since the true average causal effect is 1. Thus, we have a substantial bias induced by a relatively small association between potential outcomes and the propensity to be compromised. The logic carries through to larger experiments and more complicated examples. For instance, one can construct illustrations in which the average estimate of the average causal effect is negative, even though the individual causal effect is non-negative for every unit and positive for many units.

Notice that dropping pairs in which one pair is subject to attrition also does not recover an unbiased estimate of the causal effect for the first, second, and fourth units—that is, the units not subject to attrition. For these units, the average causal effect is $\frac{8}{3} - \frac{6}{3} = \frac{2}{3}$, so the average estimate of $\frac{1}{8}$ is quite a bit off. Whether the average causal effect defined only for non-attriters is an interesting quantity is a different, perhaps substantive question. My point here is simply that randomizing in pairs, and dropping pairs in which there is attrition, does not recover this effect, in this example and many others.

One can construct additional examples, in which there is a perfect correlation between blocked covariates and the propensity to be subject to attrition if assigned to treatment; in such examples, the strategy of dropping pairs may give unbiased estimates of the average causal effect, defined only for units not subject to attrition. In practice, however, this strategy will depend on a strong assumption about unobservables: we have to assume that all units with the same values of the blocked covariate respond similarly to treatment assignment. Yet, with just one randomization, we do not get to observe the counterfactual response to treatment assignment of control units, that is, whether they would stay in the study if assigned to treatment. So this supposition is not amenable to empirical evaluation.

Returning to the example above, notice that these problems go away if there is no attrition. Indeed, if we include all of the outcomes in the calculations, the average of the four estimates equals 1, just as it should—because without attrition, the difference of means is an unbiased estimator for the average causal effect.

Blocking Increases Efficiency; It Does Not Reduce Bias

It may be worth making one final clarifying point about the role of bias in small, unblocked experiments. As Moore (2010) points out, one of the advantages of blocking is that it can reduce the chances that the treatment and control groups are

unbalanced, at least on the blocked covariates. This is especially useful in small experiments, where the luck of the draw implies that there may be substantial imbalances across treatment and control groups on measured covariates.

However, random imbalances that arise between treatment and control groups do not lead to confounding, in the statistical sense—contrary to Moore's (2010) claim that blocking "reduces the bias in causal estimates that comes from comparing a treatment group [with covariate values] of 2, 2, 3 with a control group of 3, 4, 4." It is true that in the data, on any particular draw, covariates such as past voting history may be associated empirically with (realized) treatment assignment. Yet, bias is not the issue here: randomization ensures balance on pre-treatment covariates, in expectation. The issue is instead sampling error—that is, the random imbalances induced by random sampling of the treatment and control groups from the experimental study group—which is a different concept.

In sum, blocking can increase statistical efficiency by balancing the treatment and control groups on blocked covariates, but this is not a means of reducing bias.

Conclusion

Attrition can be quite destructive for experimental inference, when the propensity to be compromised or lost to follow-up is associated with potential outcomes. It would therefore be useful to have a way to adjust this problem away.

Yet, it appears that blocking, and dropping pairs in which one of the pairs is compromised, does not get us around this thorny problem. Blocking can be an effective tool to increase the precision of treatment effect estimators—but whether it reduces bias depends on substantive assumptions that are not justified by randomization or other aspects of the experimental design. Other strategies—for example, investing resources in tracking down units that are lost to follow up, or just a random sample of those units (Geng et al. 2008)—may offer better alternatives for reducing attrition bias.

References

Geng, Elvin H., Nneka Emenyonu, Mwebesa Bosco Bwana, David V. Glidden, and Jeffrey N. Martin. 2008. "Sampling-Based Approach to Determining Outcomes of Patients Lost to Follow-Up in Antiretroviral Therapy Scale-Up Programs in Africa." *Journal of the American Medical Association* 300(5): 506-507.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha Mara Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila and Héctor Hernández Llamas. 2007. "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): 479-509.

Moore, Ryan. 2010. "Blocking Political Science Experiments: Why, How, and Then What?" *The Experimental Political Scientist: Newsletter of the APSA Experimental Section* 1(1): 3-5.

Blocking, Balance Testing, and Analysis with `Optmatch` and `RItools`[8]

Mark M. Fredrickson
Department of Political Science, University of Illinois at Urbana Champaign
mark.m.fredrickson@gmail.com

Often, when I write about the `R` packages `Optmatch` and `RItools` , I am showing how to use them on observational data. For example, how to match observational data and test to see if the data appear balanced, as they would if they were experimentally generated. These same techniques apply to the design and analysis of experimental research. Pre-randomization blocking is matching. Reviewers and readers want demonstrations that treatment conditions are balanced on covariates. Analyzing outcomes should respect the original design, including blocking.

This paper is a "how-to" guide for using `Optmatch` and `RItools` to implement these goals.[9] It is written for the researcher who thinks blocking and randomization based inference are the right tools for experimental research and would like to see some example `R` code step-by-step. It is also for the researcher who is interested in adopting these techniques, but

[8]I thank Jake Bowers and Ben B. Hansen for their help throughout this paper.
[9]For more on the theory of blocking, see Wu and Hamada (2009) and Bowers (2010). For more on randomization inference, see Rosenbaum (2010).

is concerned about added complexity. Hopefully, this tutorial demonstrates how easy it is to block and analyze experimental data. After creating some data, I demonstrate how to create blocks of similar units, test for balance on covariates for blocked and unblocked data, and test hypotheses of treatment effects. If you wish to extract this code or make changes, the original source documents can be found at my GitHub page.

Data

Let us begin by creating some data in the style of the potential outcomes framework (Holland 1986). Let U be all meaningful covariates related to the outcomes Y_c and Y_t. We observe $X \subset U$, but do not observe $W \subset U$. The covariates are a mix of discrete and continuous random variables.

```
> n <- 100
> x1 <- rbinom(n, 10, 0.25)
> x2 <- rbinom(n, 1, 0.6)
> x3 <- rnorm(n, 50, 10)
> x4 <- rnorm(n, 0, 1)
> x5 <- runif(n, 0, 100)
> w1 <- rnorm(n, 0, 1)
> w2 <- rbinom(n, 1, 0.1)
> w3 <- runif(n, 0, 1)
> X <- data.frame(x1, x2, x3, x4, x5)
> W <- data.frame(w1, w2, w3)
```

The outcome Y is a continuous measure that is a function of the covariates and the treatment indicator. We first create Y_c from the covariates, and Y_t is simply $Y_c + \tau$, where τ is the treatment effect. The function relating the covariates to the outcome is arbitrary and only influences the analysis through the difference in potential outcomes.

```
> tau <- 10
> yc <- 0.25 * x1 + 4 * x2 + exp(x4) + x5 + 10 * w1 * w2 - w3 *
+     x3
> yt <- yc + tau
```

Blocking and Randomization

To implement blocking, we use the matching procedures in the `Optmatch` package for `R`. `Optmatch` implements a procedure known as "optimal full matching" that minimizes the average distance between matched sets (Hansen and Klopfer 2006). `Optmatch` was designed with observational studies in mind, where the researcher has discovered "treatment" and "control" groups. `Optmatch` will then find matches between similar treated and control units. This strategy is known as "bipartite matching." For more on matching (and using `Optmatch` in an observational study) see Rosenbaum 2010.[10]

In our situation, we do not have an existing randomization vector for our data, but we still wish to create similar subsets of our data. Therefore we need to create the two partitions of the data that `Optmatch` will use. The most straightforward way to create the splitting vector is to do so randomly. Just as with random assignment, in expectation these two groups should not systematically different, allowing us to find a useful blocking scheme.

```
> s <- vector("logical", n)
> s[sample.int(n, n/2)] <- T
```

To create the blocks, we use the `pairmatch` function.[11] We need to specify a distance matrix between observations,

[10] While the optimality guarantees of `Optmatch` are attractive, there are many alternative methods of blocking. See for example the `blockTools` package and a walk through Moore 2010 in the previous issue of this newsletter.

[11] `pairmatch` will create matches with one observation from each random set. `Optmatch` allows tuning the number of observations allowed from each random set. See the documentation `fullmatch` for more details.

and we can use the convenience function `mdist` to create a distance matrix based on the Mahalanobis distance between observations.[12]

```
> blocks.all <- pairmatch(mdist(s ~ x1 + x2 + x3 + x4 + x5, data = cbind(s,
+     X)))
```

For reasons of convenience or theoretical importance, we may wish to privilege certain variables and force the matching within levels of those variables. For example, if units are clustered within a geographic unit — cities within a state — we can limit matches to within the state. This is also a useful technique when matching large numbers of subjects (see my website for more details on speeding up the matching process). To limit matches within blocks, we specify a factor indicating unit membership. In our case, let use just match within the binary variable x_2. Prior to doing so, we will create a new split that places 50% of each treatment level in the partitions.

```
> count.x2.1 <- sum(x2)
> X.ordered <- X[order(x2), ]
> s.x2.0 <- sample.int((n - count.x2.1), (n - count.x2.1)/2)
> s.x2.1 <- sample.int(count.x2.1, count.x2.1/2)
> s.x2 <- vector("logical", n)
> s.x2[c(s.x2.0, s.x2.1 + (n - count.x2.1))] <- T
> blocks.x2 <- pairmatch(mdist(s ~ x1 + x3 + x4 + x5 | x2, data = cbind(s = s.x2,
+     X.ordered)))
```

For simplicity, we will continue with the single stratum blocking, but splitting up matching problems into smaller blocks is a very useful technique to have at your disposal.

Once we have blocks, we can then randomize within the blocks. As we used a pair-matching strategy, we will randomize to two treatment levels, call them "treatment" and "control." Since each observation is matched to one other we have 50 blocks with two units each. For each block, we can flip a coin and assign either the first or second unit to the treatment condition.

```
> tmp <- rbinom(n/2, 1, p = 0.5)
> z <- vector("logical", n)
> for (i in 1:(n/2)) {
+     if (tmp[i] == 1) {
+         z[i * 2 - 1] <- T
+     }
+     else {
+         z[i * 2] <- T
+     }
+ }
```

As our last manipulation to the data, create a variable that is the observed outcome (Y_c if $z = 0$ and Y_t if $z = 1$).

```
> all.data <- cbind(X, z, b = blocks.all)
> all.data$y <- ifelse(z, yt, yc)
```

Balance Testing

In expectation, randomization balances covariates, observed and unobserved. In practice, a particular randomization may place more men in treatment or more large towns in control. Experimental studies frequently provide a table of

[12] `mdist` accepts several different methods of specifying distances between observations, including a method that takes arbitrary functions. One could use this method specify distance metrics that are more appropriate to your data if you have strong theory relating the covariates to the potential outcomes. See the documentation for `mdist` for more information and examples.

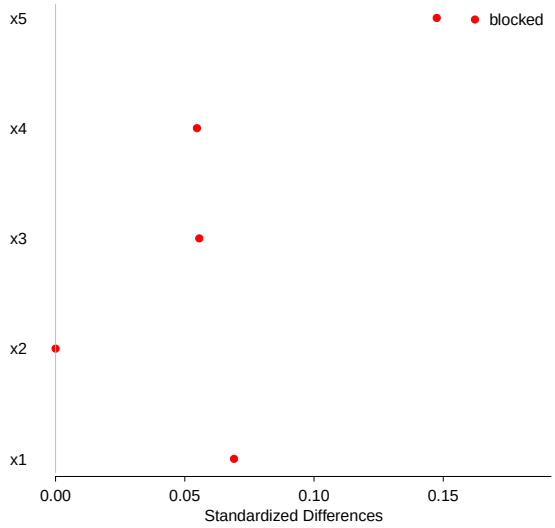

Figure 1: Standardized differences of means for simulated data.

covariates, showing descriptive statistics for treatment and control, and perhaps a set of t-tests testing the null hypothesis that the treatment and control units do not differ on each covariate.

Just as we might prefer an omnibus F-test to a series of t-tests in a regression, we might also prefer an omnibus test of balance. Hansen and Bowers (2008) provide just such a test. Instead of checking each covariate individually, the test compares all differences of means with a test statistic that converges to a χ^2 distribution. This test is implemented in the function `xBalance` in the package `RItools` .

The tool works with blocked or unblocked data. Using the blocks and treatment created in the previous section, Table 1 provides a summary of the balance testing. With a large p-value, we see little evidence that the data are not balanced (at least when considering means).

```
> observed.balance <- xBalance(z ~ x1 + x2 + x3 + x4 + x5, data = all.data,
+     strata = data.frame(blocked = all.data$b), report = "all")
```

	chisquare	df	p.value
blocked	1.41	5.00	0.92

Table 1: Omnibus balance test of balance. The null hypothesis is that the data are balanced.

We might also be interested in the balance of individual covariates. Figure 1 shows individual differences of mean (standardized) for the blocked balance assessments.[13] None of the differences are very large (all less than a quarter standard deviation's difference). `xBalance` will report per-covariate tests of balance as well, but the with the results of the omnibus test, we need not consider these.

Outcome Analysis

Figure 2 shows the treatment and control groups plotted on the outcome Y. We see a large degree of overlap in the distributions. While the mean difference of the two groups properly captures the simulated treatment effect $\tau = 10$, there is a good deal of overlap between the two groups. What is the probability that difference is simply due to chance?

[13]This is the default plot method of `xBalance` objects.

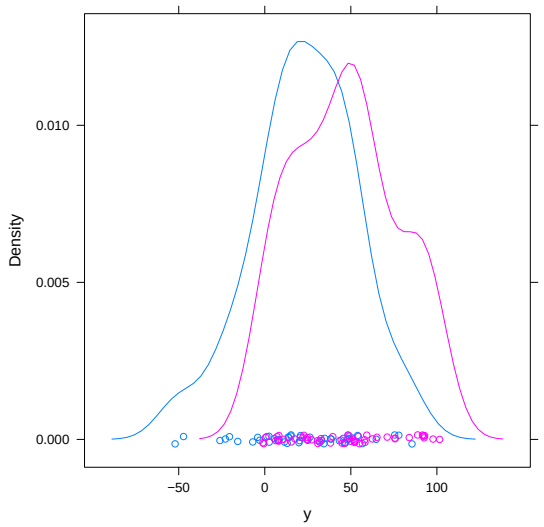

Figure 2: Outcome plot of treatment and control units

If we follow Fisher's advice that randomization is the "reasoned basis for inference", we will use a test that respects the randomization in computing the likelihood of observing the differences we see between our treatment and control groups. Two common randomization methods in `R` for computing a confidence interval for τ, the additive effect of treatment, are the `wilcox.test` in the `stats` package and `oneway_test` in the `coin` package. Sadly, neither of these functions simultaneously support blocking and confidence intervals. Fortunately, we can adapt `xBalance` to provide block aware point estimates and confidence intervals, using a Normal approximation to the randomization distribution. First we need to write a function captures the range of values we wish to test as a confidence interval.[14]

```
> f <- function(low, high, step) {
+     adjustments <- seq(low, high, step)
+     ys <- paste("I(y - z * ", adjustments, ")")
+     as.formula(paste("z ~", paste(ys, collapse = "+")))
+ }
```

For the data plotted in Figure 2, we can see that most of the mass is between -50 and 50, so we can test that range of hypotheses. We ask `xBalance` to use our blocking factor and return p-values for each of the tested hypotheses. In effect, we are inverting a series of hypothesis tests to generate a confidence interval.

```
> analysis <- xBalance(f(-50, 50, 0.5), data = all.data, strata = all.data$b,
+     report = c("z.scores", "p.values"))$results[, , 1][, 2]
> analysis.ci <- analysis[analysis > 0.05]
> analysis.ci.bounds <- c(analysis.ci[1], analysis.ci[length(analysis.ci)])
> print(analysis.ci.bounds)
  I(y - z * 6) I(y - z * 31.5)
    0.05140110      0.05710257
```

While the 95% confidence interval is rather wide at [6, 31.5], it properly rejects the null hypothesis of zero effect. Given the size of the true effect compared to the variance of the data, this interval is actually quite tight. It is unlikely that an unblocked design would find any difference.[15]

For a point estimate, we can use `xBalance` more directly.

[14] We are working at automating this process in future versions of `RItools`.
[15] While not reported here, a `wilcox.test` on an unblocked randomizations was unable to reject zero effect.

```
> xBalance(z ~ y, data = all.data, strata = all.data$b, report = "adj.mean.diffs")
     strata    strat
     stat   adj.diff
vars
y           18.93530 **
```

While larger than the true effect, given the variance, the point estimate is not that far off.

Conclusion

In the introduction I noted that the these tools, while often used for observational research, are equally at home for experimental work. I will now invert this statement to point out that the techniques in this tutorial are also useful for observational research. A matching analysis attempts to block observations on observable covariates (perhaps reduced to a propensity score). Testing balance is key to knowing if the matching achieved its goals. If we then allow the assumption that the data are "as-if" randomized, the proper analysis draws upon the randomization and blocking. `Optmatch` and `RItools` provide a set of tools that can be used for many types of analysis, both experimental and observational.

References

Bowers, J. (2010). Making effects manifest in randomized experiments. In Green, D. P., Kuklinski, J., and Druckman, J., editors, *Handbook of Experimental Political Science*. Cambridge University Press.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23:219.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3).

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Moore, R. T. (2010). Blocking political science experiments: Why, how, and then what? *The Experimental Political Scientist*, 1(1):3 – 5.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer, New York.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, analysis, and optimization*. Wiley and Sons, 2nd edition.

Section News and Announcements

- Update from APSA panel chair Kevin Arceneaux

APSA recently announced the preliminary program for the 2011 meeting (see below), and it has been a privilege to serve as program chair for the newly minted Experimental Research Division. As a healthy sign of the growing interest in experimental research, we received 115 paper proposals and 4 panel proposals on a diverse array of topics. Because this is our inaugural year, the Association allotted us only three panels. Fortunately, it was possible to organize six panels by arranging co-sponsorships with other divisions. In particular, the Formal Methods, Political Psychology, Political Methodology, Comparative Politics, and Elections and Voting Divisions agreed to partner with us on panels that feature cutting edge experimental research. I am also happy to report that a seventh panel, showcasing field experimental research on the organization of political rights across three continents, was selected by the APSA Program Committee for a coveted spot on the program as a Theme Panel.

Unfortunately, I was not able to find a home for many strong proposals. Although the large number of submissions is an encouraging indicator that some of the discipline's most promising research is employing experimental methods, it will probably be little comfort for those who could not be included in the program. The only remedy here is for all of us to make a strong showing in Seattle at the Division's panels and, in so doing, earn a larger slice of the pie at the 2012 meeting.

APSA Experimental Section Officers

- President: Jamie Druckman (2011)
- President-elect: Becky Morton (2011)
- At-large Council: Don Green (2011-2012), Mike Tomz (2010-2011), Lynn Vavreck (2010-2011)
- Treasurer: Kevin Estering (2010-2011)
- Secretary: Costas Panagopoulos (2010-2011)
- Newsletter editor: Dustin Tingley (2010-2011, renewable and appointed by President)

We are pleased to announce awards and policy committees. Awards are made possible through joining the APSA section.

- Best Dissertation in prior calendar year: Sean Gailmard (chair), Bethany Albertson, Nick Valentino, Shana Gadarian
- Best Paper at prior APSA: Josh Tucker (Chair), Rose McDermott, James Gibson, Eric Dickson
- Best Book: Ted Brader (chair), Susan Hyde, Maccartan Humphreys, Ismail White
- Junior scholars committee: Yanna Krupnikov (Chair), Rick Matland, Adam Levine, Chris Larimer, Samara Klar.
- Experimental standards committee: Alan Gerber (Chair), Kevin Arceneaux, Tom Palfrey, Cheryl Boudreau, Sunshine Hillygus, Conor Dowling.
- New experimental political science journal: John Geer (chair), Rose McDermott, Don Green, Rick Wilson, Scott Gartner, Dan Posner.

Recent Events

- **Fourth Annual NYU-CESS Conference on Experimental Political Science**

We are pleased to report that the NYU-CESS Conference on Experimental Political Science continues to go strong. The 4th Annual Conference was held March 4-5, 2011, and featured 15 paper presented and approximately 125 conference participants. The conference aims to bring together political scientists united by their use of experiments in their research, but without the normal sub-field boundaries that sometimes divide us. This year was no exception, with papers featuring survey experiments, lab experiments, field experiments, and experimental methodology on the program. Participants came from universities around the country and throughout Europe, and the conference was also bolstered by an excellent group of discussants. This was the first time we had solicited discussants though the application process as well, and this worked so well that we intend to continue it in the future. Additionally, this year's program included a themed panel Friday afternoon on experiments related to corruption. This also worked well, so we will certainly be open to considering another themed panel next year.

- **First Southern Political Science mini-conference on experiments**

January 5-8, 2011, New Orleans, Sponsored by the APSA Experimental Research Organized Section this mini-conference brought together papers by Rebecca Morton and Jean-Robert Tyran, Kevin Arceneaux and Martin Johnson, William Minozzi and Jonathan Woon, Christopher Mann, Rick Wilson and Catherine Eckel, Dan Myers and Dustin Tingley, Ngoc Phan, Cindy Rugeley and Gregg Murray, Laura Paler, Noam Lupu, and Brad LeVeck. The conferences promises to be exciting!

- **Vanderbilt Lab Experiment Meeting**

From May 4-May 6, Vanderbilt University hosted a small Conference on Laboratory Experiments in Political Science. About twenty researchers from over a dozen universities met for two purposes: first, to discuss the creation of a Consortium for Laboratory Experiments in Political Science, and second, to share innovative work based on laboratory experiments. The Consortium for Laboratory Experiments in Political Science will serve as a clearinghouse for information about laboratory experimentation. It will compile descriptions of participating experimental labs, with the purpose of facilitating conversations and collaborations across labs. It will also provide a springboard for participating researchers to investigate the possibility of trading subjects and time with researchers at other experimental laboratories. Such trades would enhance the external generalizability of laboratory studies, by providing researchers with access to additional subject pools and opportunities to field replications or extensions of projects. In addition to discussing the Consortium, researchers also presented papers that focused on experimental design, multiple agent processes, implicit and automatic attitudes, cue-taking, and biology. For more information about the Consortium for Laboratory Experiments in Political Science, contact Cindy Kam: cindy.d.kam@vanderbilt.edu.

- **WCE.2011, Caltech**

The fourth annual meeting of the West Coast Experiments Conference (WCE) was held on the campus of The California Institute of Technology on Friday, May 6.

Held each year in May, the WCE is a single day meeting that focuses on new methods for experimental research. The WCE conference is organized as a methods ”workshop” than as a venue to engage in theoretical debate. Instead of standard conference presentations, presenters focus in depth on one or two methodological take away points of their experimental work. The goal is to give the audience members applied, practical advice on methods and design in a way that will help them improve their own experimental research. We always encourage anyone with an interest in experiments to attend; graduate students are especially welcome, as well as those who are new to experimental research.

We organized this years conference around two broad themes, one for the morning and one for the afternoon. The morning panels were focused largely on design and analysis for causal inference. At the first panel, Susan Hyde (Yale Political Science) and Craig McIntosh (UCSD School of International Relations and Pacific Studies) each presented on field experimental research conducted in developing countries, Nicaragua for Hyde and Malawi for McIntosh, settings that often present great challenges for designing and implementing studies that can identify causal effects. At the second panel, Kosuke Imai (Princeton Politics) presented a new statistical method for analyzing data from an endorsement experiment to identify support for militant groups in Pakistan.

At lunch Thomas Palfrey (Caltech Economics and Political Science) gave a keynote speech summarizing research he has conducted over decades in various settings such as elections and juries, showing where Nash equilibrium predictions are robustly supported across experiments, and where they are not. Palfreys bottom line is that Nash expectations can explain a considerable amount of political behavior, but that there seems to remain a residual of behavior that requires other psychological explanations such as heuristics or risk neglect.

Palfreys keynote speech nicely foreshadowed the afternoon panels, which were designed to offer a contrast between experimental approaches adopted by those working in the economic tradition and those working in the psychological tradition. In the first afternoon panel, Becky Morton (NYU Politics) and Erik Snowberg (Caltech Political Economy) represented the economics tradition. Morton presented lab experiments designed to test whether participants vote to express preference for a candidate or whether they vote with an eye to how a legislative assembly is finally constituted. Snowberg presented a theoretical paper that framed subjects participation in an experiment as a principal-agent problem, a framework that can identify not only treatment effects but also subjects beliefs about the effectiveness of the treatment itself.

The second afternoon panel focused on psychological approaches. Shana Gadarian (UC-Berkeley Public Health) outlined new methods for eliciting anxiety in lab experiments, methods that can be embedded in experiments to test the effects of anxiety on political participation, information gathering, and attitudes. Wendy Wood (USC Psychology) presented a study that demonstrated the habitual nature of voting within a laboratory setting.

There were about 60 participants in all at the conference. The conference was a great success: the presentations were informative and the audience was quite active in the discussions. Caltech was a gracious host and served as a magnificent venue for the conference.

We expect to begin organizing WCE.2012 in the coming months. We will announce plans for next years conference in a subsequent newsletter.

Kevin Esterling, Nick Weller, Mat McCubbins, Jonathan Katz

Upcoming Events

- **Fifth Annual NYU-CESS Conference on Experimental Political Science call for papers**

Please mark your calendars for next year's conference, which will take place on March 2-3, 2012. Joshua Tucker and Eric Dickson will be co-directing the conference, and we expect to post a call for papers early in the fall. Thanks to the generosity of NYU's Department of Politics and the NYU's Center for Experimental Social Science we continue to be able to offer the conference without a registration fee and to provide free food throughout the weekend, so we hope to see as many of you there as possible!

- **2011 APSA Annual Meeting**

Lots of great section sponsored and co-sponsored panels! Room assignments pending.

Section 51: Experimental Research
51-1 Formal Models and Experiments: the New Frontiers, Thursday, Sep 1, 2011, 10:15 AM-12:00 PM
51-2 How People Think and Deliberate About Politics, Friday, Sep 2, 2011, 4:15 PM-6:00 PM
51-3 The Impact of Transparency on Government Performance: Comparative Evidence from Field Experiments, Sunday, Sep 4, 2011, 8:00 AM-9:45 AM
51-4 Experiments on Cue-taking, Myopia, and Rationality, Thursday, Sep 1, 2011, 2:00 PM-3:45 PM
51-5 Experiments on Social Pressure and Mobilization, Thursday, Sep 1, 2011, 8:00 AM-9:45 AM
51-6 Methodological Issues in Randomized Experiments, Saturday, Sep 3, 2011, 4:15 PM-6:00 PM
51-7 Theme Panel: Organizing for Rights in Africa, Asia, and the Southwest: Field Experiments in Heterogeneous Communities, Friday, Sep 2, 2011, 8:00 AM-9:45 AM

Back to Contents