



The Experimental Political Scientist



In this issue

- **Berinsky, Quek, Sances: mTurk**
- **Mintz: Decision Process Tracing**
- **Carlin and Love: Behavioral Knowledge Base**
- **McClendon: Public Officials and Field Experiments**
- **Peyton: Ethics and Field Experiments**
- **Titunik: Review of New Gerber/Green Book**
- **Announcements**

From the Editor

Welcome to the next issue of the *Experimental Political Scientist*. Leading off are discussions of experimental “technologies” and associated research programs: mTurk for recruiting subjects throughout the world and a platform for decision process tracing. Next our colleagues take on a discussion of the role behavioral game experiments that arose from a recent conference. Next we feature several contributions about field experiments. We wrap up with announcements from across our community. Thanks to all the contributors. Special thanks to some new “field” editors: Dan Myers, Jonathan Renshon, and Kristin Michelitch. PLEASE send contributions (or ideas) for the fall newsletter, deadline October 15th. 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>. Previous issues available. Send submissions to the [editor](#).

Journal Coming Soon!

In December 2011, the Section voted overwhelmingly to establish a new journal, the *Journal of Experimental Political Science*. We will serve as co-editors with the assistance of an active group of Associate Editors: James Druckman, Northwestern, Ray Duch, Oxford, Thad Dunning, Yale, Sean Gailmard, UC Berkeley, Donald Green, Columbia, Leonie Huddy, Stony Brook, Kosuke Imai, Princeton, Bernhard Kittel, Vienna, Esteban Klor, Hebrew U. of Jerusalem, Rose McDermott, Brown, Costas Panagopoulos, Fordham, Jeremy Weinstein, Stanford, and Liz Zechmeister, Vanderbilt. A contract has been offered and is currently being finalized. We hope to begin accepting papers in late summer or early fall 2012. We will notify members of the Section as soon as we are able to begin receiving papers through the email list; at that time we will also provide instructions for submissions. We anticipate the first issue of the journal to appear in print in 2014 (although papers accepted before then will appear online and be available as soon as they are ready). The journal will not be a success, however, unless we receive high quality submissions and we encourage all section members to plan on submitting papers (and helping out with the review process)!

Rebecca Morton and Joshua Tucker, NYU

Conducting Online Experiments on Mechanical Turk

Adam Berinsky, Kai Quek, and Michael Sances

Massachusetts Institute of Technology

berinsky@mit.edu

quek@mit.edu

mwsances@mit.edu

Experiments can contribute important knowledge to political science, but they are often costly to implement. Amazon.com's Mechanical Turk allows researchers to conduct internet-based experiments at minimal cost. In this brief article, we highlight the promise and pitfalls of Mechanical Turk.

What is Mechanical Turk?

Mechanical Turk (MTurk) is an Internet service provided by Amazon.com where prospective workers, who seek payment for the performance of short tasks, may be matched with prospective employers, who seek workers to perform these tasks. These tasks are known on MTurk as HIT's, or Human Intelligence Tasks. MTurk is particularly useful in the recruitment of experimental subjects. Economists, psychologists, and political scientists are making increasing use of MTurk to recruit subjects, paying anywhere between 0.10 and 0.75 for the completion of internet surveys (e.g., Berinsky, Huber, and Lenz 2012; Chandler and Kapelner 2010; Chen and Horton 2010; Paolacci, Chandler and Ipeirotis 2010; see also Bohannon 2011). The logic of using MTurk is compelling: recruitment in laboratory experiments usually costs between \$5 to \$30 per subject. MTurk is many times cheaper compared to other alternatives.

How does MTurk work?

Using MTurk is straightforward. First, the researcher will set up a Requester account on www.mturk.com and add money to that account. Thereafter, the researcher can walk through a few guided steps on the MTurk interface to create a new HIT. The HIT describes the task to prospective workers, as well as the payment for completion. The researcher can set specific requirements for the HIT, such as the respondent's country of residence, the approval rate by other Requesters for prior HITs, and the number of times an individual respondent can perform the task. For political scientists, the task can be a survey that can be hosted either on MTurk itself or on an external survey platform such as SurveyMonkey or Qualtrics. In the latter case, the respondent should receive a link on the MTurk webpage to the external webpage where the survey is hosted; at the end of the survey the respondent can be given an individualized code to be entered back into MTurk for payment purposes. MTurk also provides the option for variable bonus payments to particular respondents identified by the researcher.

Does MTurk work well?

Berinsky, Huber and Lenz (2012) (henceforth: BHL 2012) found that based on observable characteristics, the U.S.-based MTurk sample is more representative of the national population than convenience samples of college students, but less representative than national probability samples. BHL (2012) showed that the MTurk sample is demographically very similar to the sample obtained in the

unweighted American National Election 200809 Panel Study (ANESP), a high-quality survey study which recruited respondents through a random-digit-dial method for a 21-wave Internet-based panel survey. However, there are important differences between the MTurk and ANESP samples in political ideology, the level of political knowledge and partisan self-identification. MTurk respondents tend to be slightly more Democratic than ANESP respondents, as well as more liberal and knowledgeable about politics. BHL (2012) also found that the MTurk sample is younger on average compared to respondents in the Current Population Survey, and consequentially the MTurk sample also differs on age-related characteristics such as marital status, homeownership, and religious preference. Finally, BHL (2012) successfully replicated several survey experiments from psychology and political science.

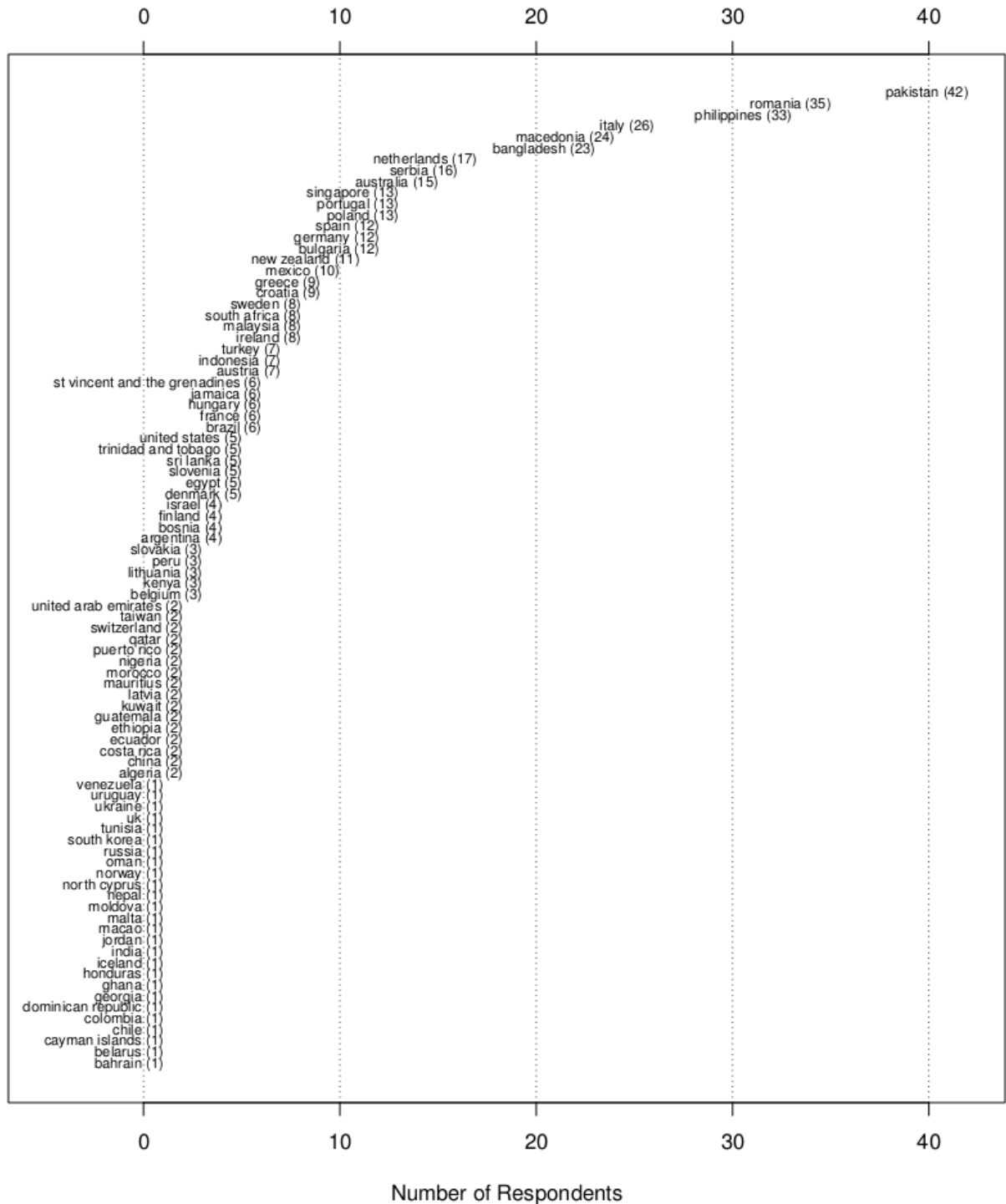
While MTurk is a powerful tool for conducting low-cost political science experiments, it is currently unclear how useful it is for recruitment beyond the U.S. A piece of folk wisdom among experimentalists who use MTurk suggests that the pool contains a large proportion of Indian workers. Lawson et al. (2010) leverage this fact by employing samples of both American and Indian MTurk workers to rate faces of Mexican and Brazilian political candidates. Of their 50 Indian raters across their three tasks, 67-70% were male, the average age was 28, and more than 80% had college degrees; no other demographic information was collected. Lawson et al. (2010, 569 note 42) also report that they had to disqualify many potential Indian raters due to unusual behavior, such as giving nonsensical responses or clicking the same choice for all questions.

Beyond the folk wisdom, there are the official numbers from Amazon.com. In June 2011, Amazon.com put the number of workers at 500,000 and the number of countries at 190 (Varia 2011). Even with a potentially rich pool, however, there are some important barriers to employing foreign workers on Mechanical Turk. For instance, MTurk initially only paid workers in U.S. dollars (subjects could also choose to receive credits on Amazon.com), effectively freezing out any non-U.S. respondents; the policy was later modified to allow for payment in Indian rupees as well (Ipeirotis 2010).

A handful of existing studies have begun to quantify the geographic origins of the MTurk population, and have largely supported Amazon's claims of an internationally diverse pool. However, these studies do not give a breakdown of what countries are included in the residual category, beyond the U.S. and India. The exception is Ross et al. (2010), who report 11% of respondents from other countries in their sample of 573: Canada (3%), the Philippines (1%), the UK (1%), Pakistan (0.5%), Romania (0.5%), other countries (5%), and an unspecified proportion of respondents from Australia and Ukraine.

To our knowledge, no studies have been conducted to ascertain the geographic reach of MTurk and to characterize the respondents coming from non-U.S. countries. Hence, we conducted two studies where we interviewed subjects from MTurk outside North America and India (Berinsky, Quek and Sances 2012; henceforth: BQS 2012). Figure 1 illustrates our finding on the respondent distribution in the global MTurk sample, based on our Study 1 (N = 528).

BQS (2012) also measured demographic information such as nationality, country of residence, age, gender, education level, marital status and socioeconomic status. We also measured the respondent's comprehension of the English language as well as the respondent's attention to the survey questions. Finally, like BHL (2012) we replicated several experiments in psychology and political science.



BQS (2012) found that most subjects in the global MT pool are male respondents between 20-35 years old; college-educated; white-collar workers (if employed); proficient in the English language; and

attentive over the course of the survey. Based on a non-U.S. MTurk sample, BQS (2012) conducted successful replications of the classic framing experiment by Tversky and Kahneman (1981); Kam and Simas’s (2010) experiment in American politics; and Tomz’s (2007) experiment in international relations. Table 1 compares the main results from Kam and Simas (2010) with the replicated results in BHL (2012) based on a U.S. MTurk sample and the results from the non-U.S. MTurk sample in Study 2 of BQS (2012). In all three experiments, support for the risky policy option increased for subjects assigned with the mortality frame. The coefficient of the treatment variable (effect of the mortality frame) in our replication (1.066) is almost the same as that in Kam and Simas’s (2010) study (1.068) and that in BHLs (2012) replication (1.180). Our study also replicated the null result from Kam and Simas of no significant interaction between the mortality frame and the risk measure.

	Kam and Simas (2010)		BHL (2012)		BQS (2012)	
Mortality Frame	1.068 (0.097)	1.058 (0.294)	1.180 (0.10)	1.410 (0.31)	1.066 (0.199)	0.986 (0.703)
Risk Acceptance	0.521 (0.306)	0.507 (0.481)	0.760 (0.29)	0.990 (0.42)	1.660 (0.721)	1.577 (1.006)
Mortality × Risk		0.023 (0.624)		-0.450 (0.58)		0.171 (1.442)
Intercept	-0.706 (0.155)	-0.700 (0.227)	-1.060 (-0.170)	-1.190 (-0.230)	-1.361 (0.376)	-1.321 (0.506)
lnL	-453.185	-453.184	-409.740	-409.439	-125.3	-125.3
$p > \chi^2$	0.000	0.000	0.000	0.000	0.000	0.000
N	752	752	699	699	181	181
Sample	Knowledge Networks		MTurk US		MTurk Global	

Where MTurk might not work well

MTurk may or may not be an appropriate experimental platform depending on the type of experiment designed and the specific research question asked by the experimenter. As it is, MTurk is more suited for internet-based survey experiments than incentivized experiments with real-time interactions, or laboratory experiments where the environment must be tightly controlled and the subjects directly observed. Furthermore, there are atypical characteristics in the MTurk sample that may bias the estimated treatment effects under certain conditions. For instance, MTurk subjects in the U.S. seem to be relatively young, Democratic, liberal, and knowledgeable about politics (BHL 2012). For developing countries, MTurk is very likely to be over-sampling the sub-population that is more internet-savvy, more proficient in English, and more exposed to Western influence than the average

citizen (BQS 2012). These atypical characteristics may threaten the validity of the experimental treatments if the treatments are sensitive to these characteristics. As BHL (2012, 11) concluded, if we treat the MTurk as a means for conducting internally valid experiments, instead of a representative sample, the MTurk respondent pool is very attractive. At the same time, if one is interested in estimating treatment effects that may differ due to any of the factors for which the MTurk sample is less representative, then the MTurk sample may yield estimates that are inaccurate for the larger population.

With a clear understanding of its strengths and limitations, a political scientist may find MTurk a highly efficient tool for pilot probes of research ideas and for low-cost implementation of large-N internet-based experiments. Furthermore, as access to the Internet expands across the globe, drawing diverse international samples through MTurk will become a real possibility. On the whole, we believe that MTurk can be very useful to political scientists as long as it is used appropriately.

References

- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Using Mechanical Turk As a Subject Recruitment Tool for Experimental Research." Forthcoming in *Political Analysis*.
- Berinsky, Adam J., Kai Quek, and Michael W. Sances. 2012. "Using Mechanical Turk As a Subject Pool in Developing Countries." Paper prepared for the Midwest Political Science Association Annual Conference. Chicago: April 2012.
- Bohannon, John. 2011. "Social Science for Pennies." *Science* 334 (6054): 307.
- Chandler, Dana, and Adam Kapelner. 2010. "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets." University of Chicago Mimeo.
- Chen, Daniel L., and John J. Horton. 2010. "The Wages of Pay Cuts: Evidence from a Field Experiment." Harvard University Mimeo.
- Kam, Cindy D., and Elizabeth N. Simas. 2010. "Risk Orientations and Policy Frames." *Journal of Politics* 72(2): 381-396.
- Lawson, C., Gabriel S. Lenz, Mike Myers, and Andy Baker. 2010. "Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies." *World Politics* 62(4): 561-93.
- Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5): 411-19.
- Ross, J., Irani, I., Silberman, M. Six, Zaldivar, A., and Tomlinson, B. 2010. "Who are the Crowdworkers?: Shifting Demographics in Amazon Mechanical Turk." Paper, UC Irvine Department of Informatics.
- Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4): 821-40.
- Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481): 453-58.
- Varia, Jinesh. 2011. "Get Better Results with Amazon Mechanical Turk Masters. Amazon Web Services Blog: <http://aws.typepad.com/aws/2011/06/amazon-mechanical-turk-master-workers.html>. (Accessed on 5 April 2012)

The Pentagon Computerized Process Tracing Experiment

Alex Mintz

Dean of Lauder School of Government, Diplomacy and Strategy at IDC-Herzliya

mintz.alex@idc.ac.il

In a study conducted with high ranking military officers at the National Defense University, we have tried to identify, using computerized process tracing experiments, whether military leaders are susceptible to information processing biases in decision making.

We presented high ranking military officers who participated in an NDU course on leadership, a decision scenario and a decision matrix. The decision scenario featured a decision problem-selecting among technologies for combating terrorism, whereas the decision matrix consisted of four alternatives (border crossing sensors, environmental monitors, local emergency responders, and do nothing) and three dimensions (military, economic, political).

Participants in this study accessed the Decision Board software which recorded how they arrived at a choice, their "Cognitive Algebra" and their cognitive biases.

Some of the key findings uncovered in this project, reported in Mintz, Redd, and Vedlitz (2006) and Mintz and Redd (2007), are:

- 1) Leaders are susceptible to information processing biases. Specifically, the military officers had a "preference over preference" to Do Something instead of Do Nothing;
- 2) they also locked in on one alternative instead of examining a range of alternatives;
- 3) in many instances, they ignored critical information.
- 4) they used satisficing strategies rather than maximizing processing;
- 5) they rated the military dimension as the most important factor influencing their decision and the economic dimension as the least important dimension.
- 6) Leaders make decisions in a two stage process: they eliminate alternatives which hurt them politically; they then conduct an analytic, rational analysis in an attempt to arrive at the optimal choice but only on the alternatives that "survived" the first stage of this decision process. This process is called Poliheuristic Theory (Mintz 2004a).

Svenson (1979:98) notes that the aim of process tracing is to reveal "a cognitive process, leading to a final decision or solution". Process-tracing methods focus on process and outcome validity, or the manner in which decision makers process information en route to a choice.

As Mintz, Geva, Redd, and Carnes(1997:556) state, "[process tracing's] main strength is its ability to identify specific strategies used by decision makers and to test theoretically derived implications of situational and personal variables on the decision process and its outcome." Process-tracing techniques have been used to investigate political decision-making processes, strategies, and outcomes in American politics see for example, Lau and Redlawsk 1992 and Riggle and Johnson 1996). In international relations research, experimental process-tracing examining foreign policy and national security decision making have been undertaken largely with the Decision Board platform.

The Decision Board Simulator Platform is an Internet-based computerized decision process tracer that records key features of the decision making process. A decision task is typically presented to the subjects in the form of a scenario and a decision matrix consisting of decision alternatives and relevant dimensions. The subjects choose from a set of alternatives based on information that can be accessed from the computer. The Decision Board records key features of what the subject examines prior to making his or her choice. This information is then analyzed to make inferences about process, choice, information processing biases and the decision profile of users.

The computerized process tracing methodology along with the Decision Board technology enabled our research team to also test experimentally:

- 1) The effect of dynamic versus static choice sets on information processing and choice (American Political Science Review 1997)
- 2) The effect of negative political information on processing and choice (2005)
- 3) The effect of familiar versus unfamiliar decision problem on information acquisition and choice (Journal of Conflict Resolution 2004b)
- 4) The effect of ambiguity on process and choice (Journal of Conflict Resolution 2004b)
- 5) The effect of framing and risk propensity on choice (ISA 2007, also see Frisch 2000)
- 6) The effect on choice of the appearance of a new dimension during the decision process (Geva et al 2006)
- 7) Groupthink versus Polythink (Mintz and DeRouen 2010)

“Real world” decision makers such as high ranking military officers at the US Air Force academy and at the National Defense University participated in these studies. The Decision Board has also been utilized in workshops at the United Nations, in Executive Training courses, in developing societies, and in many leading universities worldwide. It is the most popular information board in International Relations, foreign policy analysis and national security analysis. For more information on the potential use of this software, email the author at mintz.alex@idc.ac.il.

References

- Frisch, D., 1993. Reasons for Framing Effects. *Organizational Behavior and Human Decision Processes*, Volume 54: 399-429.
- Geva, N., S. Redd and A. Mintz., 2006. Evolving Vs. Static Decision Dimensions and Decision Processes: An Experimental Assessment of Poliheuristic Propositions. Unpublished paper.
- Mintz, A., 2004a. How Do Leaders Make Decisions? *Journal of Conflict Resolution*, Volume 48, No. 3. February.
- Mintz, A., 2004b. Foreign Policy Decision Making in Familiar and Unfamiliar Settings: An Experimental Study of High-ranking Military Officers. *Journal of Conflict Resolution*, Volume 48, No. 1. February.
- Mintz, A., 2005. Are Leaders Susceptible to Negative Political Advice? An Experimental Study of High-Ranking Military Officers. In A. Mintz and B. Russett (eds.) *New Directions for International Relations*. Lahman, MD: Lexington Books.
- Mintz, A. and Redd, S., 2007. Biases in Decision Making and their Influence on Counterterrorism

- Decisions: A Computerized Experiment. Paper prepared for presentation at the annual meeting of the American Political Science Association, Chicago, Illinois, August 29 -September 2.
- Mintz, A. and DeRouen K., 2010. Understanding Foreign Policy Decision Making. Cambridge University Press. (Section on Polythink).
- Mintz, A, Geva, N., Redd, S. and Carnes, A., 1997. The Effect of Dynamic and Static Choice Sets on Political Decision Making: An Analysis Utilizing the Decision Board Platform. American Political Science Review, Volume 9, No. 3. September.
- Mintz, A., Redd, S. and Vedlitz, A., 2006. Can We Generalize From Student Experiments to the Real World in Military Affairs, Political Science and International Relations? Journal of Conflict Resolution, Volume 50, No. 5. October.
- Redlawsk, D and Lau, R., 1992. How Voters Decide: A Process Tracing Study of Decision Making in Political Campaigns. Paper presented at the annual meeting of the American Political Science Association. Chicago, IL.
- Riggle, E., and Johnson, M., 1996. "Age Difference in Political Decision Making: Strategies for Evaluating Political Candidates." Political Behavior, Volume 18, No. 1. March.
- Svenson, O., 1979. Process Descriptions in Decision Making. Organizational Behavior and Human Performance, Volume 23, No. 1. February.

Building a Behavioral Knowledge Base? Challenges Facing a Multidisciplinary Research Agenda

Ryan E. Carlin
Georgia State University
rcarlin@gsu.edu
Gregory J. Love
University of Mississippi
gilove@olemiss.edu

Game-theoretic behavioral experiments are increasingly used to study social preferences in a range of disciplines, including political science. Indeed behavioral instruments from simple dictator games to elaborate multi-round common pool resources games have appeared in our disciplines leading journals. While this is a potentially sanguine development from the perspective of theory building, the presentations, debates, and comments made at a recent workshop funded by the Mellon Foundation, the Latin American Studies Association, and Georgia State University (GSU) unearthed several key challenges to knowledge accumulation confronting this multidisciplinary effort. Here we focus on three: (1) experimental environments, (2) measurement, and (3) language. For these experimental approaches to grow in prominence and offer even greater insight into social preferences and political choices, it behooves scholars to take these matters seriously.

Experimental Environments

Across the social and neural sciences, experimentation takes places in a wide array of environments. A key conclusion from the conference is that conducting experiments in lab and non-lab environments

implies a series of tradeoffs and dilemmas. For example, experimentalists seeking to contribute not only theoretically but also to practical problems face some Catch-22s. One on hand, practitioners interested in the programmatic or policy applications of lab-based experiments often complain that lab-based designs are so complex that they confuse subjects outside of the lab. On the other hand, colleagues unfamiliar with experiments regularly bemoan the same designs as far too simple. And once the results of behavioral experiments are published, scholars, subjects, and program participants themselves are often skeptical as to whether one can extract any real-world lessons from behavior in such abstract and deceptively simple games. Thus simultaneously convincing fellow scholars, practitioners, policymakers, and participants of the value of behavioral experiments is no mean feat.

The dilemmas these diverse audiences pose should influence our research designs and goals. As Rick Wilson argued, there are good reasons to leave the lab. Moving from college students to new populations with diverging social preferences embedded in intriguing contexts could yield major theoretical gains. It also probes external validity by providing substantial subject heterogeneity. And finally, beyond the lab there are greater chances for serendipity to play a role in the research. Insights into the how social preferences are formed or influence behavior often begin as intuitions formed by chance events. Yet, conversely, Wilson cited good reasons not to leave the lab as well, such as the expense, time investment, logistical problem, unforeseen difficulties, and, in some places, danger. Therefore, experimentalists are urged to test and refine their theory and instruments in the lab before deciding whether to incur the costs and burdens of conducting experiments outside it.

Finally, the choice whether or not to leave the lab must reflect ones analytic goals. The extent and nature of the empirical claims that one wishes to make should dictate the nature, location(s), and subjects of the study. As outlined in Tara Grilloss and Richard Zechausers presentation, and backed by recent work by Henrich et al. (2004, 2010), if experimentalists want to make generalizable claims about broad principles of human behavior they must marshal a tremendous amount of evidence. First, to show generalizability, their findings must hold in multiple contexts to convince others that the behavior is not bound by locality, country, or region. Second, to lay claim to a key principle of human behavior, experimentalists must construct a compelling theory whose testable implications bolster the principles breadth. These propositions can push analysts beyond the friendly confines of the lab into field labs and, potentially, towards RCT-style field experiments. Viewed in this light, constructing general knowledge about core principle human behavior can be a lonely and weighty endeavor. But viewing it as a long-run multidisciplinary goal helps lighten the load.

Measurement

What do social preference-focused behavioral games actually measure? As research using dictator, trust, and public goods games become increasingly common in the social sciences, it is critical to continually assess the validity of these measures. At the first step, researchers should clearly specify the theoretical concept they seek to measure. What exactly does one mean when discussing altruism, selfishness, trust, reciprocity, or cooperation? Secondly, does this specific definition match the incentives in the game? Do researchers and subjects understand these incentives in the same way? And might these incentives be tapping into more than one social preference (trust and risk, for example)?

Several workshop participants who do not employ these approaches emphasized the need for strong linkages between clearly defined theoretical concepts and experimental games. A closely controlled lab RCT helps ensure a causal relationship between treatment and effect. But with poorly understood measures it is unclear what, the treatment has caused. And, as Cindy Kam argued, many of the behavioral games political scientists increasingly use have their origins in other disciplines, especially economics and social psychology. When adapting these games to investigate questions central to political science it is important to consider how the original purpose of the measure, and the evidence to support its validity, maps on to how political scientists use them.

This conclusion applies *a fortiori* for a final measurement issue highlighted in the presentations of Juan Camilo Crdenas and Rick Wilson and echoed by others. Experimentalists in a variety of fields are seeking novel adaptations of the classic experimental games to better reflect the complexity of social reality. The challenge is how to move beyond the static games with a few players and a few rounds (often one), towards models with ecological factors that increasingly mirror real world scenarios *without sacrificing tractability*. One could imagine ever more dynamic and sequential games, or games with multiple resources and/or that account for spatial relationships and the role of communication. But by increasing complexity we often decrease our ability to interpret theoretically and empirically the behaviors observed in more complex experiments.

Language

A surprising but crucial aspect that came from workshop discussions was a fundamental point of potential cross-disciplinary disagreement/confusion on what constitutes an “experiment. While many political scientists reading this newsletter would likely agree that random assignment is a *sine qua non* for experiments, this belief may not be widely held in our related disciplines or even among other political scientists. Indeed, some would argue the classic game- or decision-theoretic behavioral games qualify as experiments even when the ultimate goal of such games is measurement. Strong causal claims, experimental political scientists will often insist, require random assignment to treatment and control groups. Occupying a middle ground within this debate are research designs that rely on within-subject comparisons between a baseline and treatment behaviors. Moreover, in sequential games involving two or more players all but the first player cannot be said to receive a randomly assigned treatment, as their actions are clearly endogenous to previous plays. In sum, there is little consensus within and across disciplines on whether game- or decision-theoretic behavioral experiments are merely measurement tools or methods for determining causality.

A second and related linguistic discrepancy the conference unearthed concerns “field experiments. For many political scientists, a field experiment suggests a design by which randomly assigned groups either receive a treatment or not, measurement is taken, and causality is assumed insofar as the random assignment “worked and, thus, controlled for other explanations. For many economists, a field experiment attempts to replicate a behavioral game outside the lab (lab-in-field, for many political scientists). Typically, scholars will ascribe causality to the contextual variation in which the behaviors are observed. This tension speaks directly to the aforementioned challenges of tradeoffs involved in the choice of lab versus field.

Third and finally, the term “external validity differs in meaning for experimentalists across disciplinary borders. For some, this chiefly refers to whether behavior observed in the lab matches behavior observed in the real world. Social and political psychologists, on the other hand, often establish external validity via replication across a range of (not necessarily representative) samples. Still others will want to judge external validity with respect to the population to which one would like to generalize.

These differences in definition, understanding, and the applicability of experiments in such closely related fields as political science, economics, psychology, sociology, and experimental anthropology bring the challenges of cross-discipline knowledge accumulation into sharp relief. Until social scientists agree on the meanings of these crucial concepts, researchers are well advised to make definitions explicit and transparent. In doing so, we can more clearly delineate the value of causal inferences and the utility of our approach.

Conclusion

Behavioral experiments developed to empirically test game-theoretic models hold a great deal of promise for understanding phenomena of interest to political scientists. Luckily, we can rely on the vast and growing research in other fields to inform our designs. This does not, however, let us or anyone else off the hook for the lack of clarity regarding measurement, context, validity, and inference. Going forward, we can meet the challenge of building a behavioral knowledge base by being up front and transparent about these and other issues that still challenge behavioral experimentalists rather than sweeping them under the rug. We hope this brief note can energize the debates surrounding the value and challenges of employing these methods by enjoining political scientists to take an active part in shaping this cross-discipline research agenda.

References

- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr and Herbert Gintis (eds.). 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. New York: Oxford University Press.
- Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwina Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker. 2010. Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science* 327, 1480-1484.

Ethics of Using Public Officials As Field Experiment Subjects

Gwyneth H. McClendon
Department of Politics
Princeton University
gmcclende@princeton.edu

In recent years, the number of field experiments using public officials¹ as subjects has increased.² Some of these experiments are relatively cheap and easy to replicate, and yet, the ethical guidelines for dealing with this particular class of subjects are in some ways both nascent and underdeveloped. If this line of research is to continue in a responsible manner, it is worth considering whether there are particular ethical questions raised by using elite, public servants as subjects and if so, how we might deal with those ethical questions. I am not a moral philosopher, and I do not have hard and fast answers to the issues I raise below. I am sympathetic to these types of experiments, having conducted one myself, but I am also sensitive to issues raised by their critics. I hope the following will encourage further debate.

There are at least two reasons public officials differ from citizens as experimental subjects. First, when we use public officials as subjects, the behavior we are trying to observe and affect is usually part of their routine, public duties. This may mean that the societal benefits of the knowledge gleaned from these experiments are particularly great. But it also introduces questions about *whose* resources are being used while subjects take time to participate, as well as questions about debriefing and appropriate compensation.³ Second, public officials are a limited pool, and they often control public

¹By public official, I mean a person “elected or appointed to carry out some portion of a government’s sovereign powers”(Black’s Law Dictionary).

²To give a few examples, Butler and Brookman (2011) conducted an email experiment with United States legislators as subjects that tested for racial bias in legislators’ responsiveness. Each legislator was contacted by a fictitious constituent, whose name signaled racial group membership. The subjects of Loewen and MacKenzie (2011)’s study were Canadian members of parliament and provincial legislators. These subjects were also contacted via email by fictitious constituents who asked the legislators about issues inside and outside of their jurisdiction. Loewen and Rubenson (2011) conducted an experiment with elected delegates in a Canadian Liberal Party leadership contest. Some subjects were randomly assigned to receive pieces of direct mail to persuade them to vote for a particular candidate. Humphreys’ and Weinstein’s ongoing study of transparency in Uganda looks at the effects of performance evaluations on the behavior of Members of Parliament. Working with an NGO, the authors designed an annual report card about MP activities inside and outside of parliament and disseminated the report card more heavily for some MPs than for others so that they could observe whether MPs alter their behavior in response to greater transparency. Malesky, Schuler and Tran (2011) used delegates to the Vietnamese National Assembly as their subjects and examined whether randomly selected delegates improved their performance in query sessions when websites documenting their performances were made public. I conducted a field experiment with South African local politicians that tested for racial bias in response rates (McClendon, 2012). Subjects were contacted by fictitious constituents with queries about public goods provision.

³Another possible issue is that, by affecting the enactment of public duties, researchers may also alter the outcome of large, communal events. Approaches to minimizing this ethical concern have been discussed in detail elsewhere (e.g. Wantchekon, 2003; Loewen et al., 2010), so I leave them aside here.

purse strings, including funding for universities and research. This difference introduces questions about researchers' ethical responsibilities to *other researchers* when conducting experiments with public officials. Let me consider each of these differences in turn.

Responsibilities to Subjects

Most if not all field experiments involving public official subjects focus on the execution of the officials' routine, public duties. Consider, for instance, field experiments that examine the responsiveness of elected officials to their constituents (Butler and Brookman, 2011; Loewen and MacKenzie, 2011; McClendon, 2012). The outcome of interest—whether politicians respond and are helpful to constituents—is a routine part of the subjects' public duties. If there were no deception and if participants consented to having their communications observed, the research would almost certainly be exempt from IRB review. IRB regulations regard public officials as a population that is less vulnerable than other typical populations of subjects. Appointed public officials and candidates for public office expect to be regularly scrutinized as part of their jobs. They do not have the same expectations of privacy and confidentiality as the average subject, and one might argue that the public has a right to know whether elected politicians execute their public duties appropriately and in good faith.

Yet, most, if not all, field experiments involving public official subjects do not ask for subjects' consent in advance of the research, and many use deception. Studies that test for racial, ethnic or gender bias in politicians' responsiveness, for instance, typically use fictitious constituents to contact unknowing elected representatives. In this way, they are very similar to a host of field experiments that have looked at discrimination in the labor and housing market (Pager et al., 2009; Bertrand and Mullainathan, 2004; Brown and Gay, 1985). Other studies, such as those that ask whether greater transparency changes the behavior of public officials, do not use deception, but they typically waive informed consent. In either of these cases, should public official subjects be compensated after the fact? Should they be debriefed? Should this type of research not be conducted at all?

There are good reasons for which these studies use deception and forego informed consent. A key advantage of field experiments over other types of experimentation is that they examine behavior in a real-world setting. This often means that the experimental conditions are naturally occurring or occur frequently enough without researchers' interventions such that subjects have implicitly consented to experience them as part of normal life (Humphreys, 2011). Whether researchers are involved or not, elected representatives receive emails from constituents on a daily basis; they are often scrutinized for their attendance and participation in legislative and query sessions. A key aspect of a "real-world" setting is also that subjects do not know they are being observed. Indeed, asking for informed consent is likely to destroy the realism of the experiment and to jeopardize the inferences that could be drawn. When the behavior being affected is socially distasteful (which discrimination, shirking and other poor executions of public duties are), not asking for informed consent in advance may be the only way to observe the behavior accurately.⁴ Whether public official or citizen subjects are involved, IRBs allow

⁴Some might argue that, while the waiving of informed consent or the use of deception may be the only way to arrive at reasonable inferences to some important research questions, political scientists should still not conduct such studies, leaving the "dirty" work to journalists, non-governmental and/or international organizations. I would argue, however, that political scientists often have the best training to execute these studies and that well-conducted studies are better than bad studies.

informed consent to be waived where interventions are ordinary, where answering the research question is of high social value and where obtaining informed consent in advance would damage the results of the experiment.

Similarly, most field experiments that use deception with public officials do so *only* because accurate answers to questions about officials' public duties are of high social value and because no other method would yield accurate inferences about the particular questions they want to answer. Field experiments that test for racial discrimination in the political sphere, for instance, use fictitious constituents because controlling the names and traits of the constituents and the exact content of the communication is the only method that allows researchers to detect accurately whether it is race *per se* and not some other factor driving politicians' average rates of response. These studies typically take precautions to ensure that the intervention poses as little inconvenience to the subjects as possible; and they typically use a between-subject design so that no inferences can be drawn about the biases of any one public official. These measures further minimize potential harm to the subjects, and the deception itself reveals harmful behavior that could not otherwise be detected.

Indeed, if anything, deception is slightly less ethically troubling when using public official subjects than when using other types of subjects. Lies are an endemic part of the political sphere, in a way that is not true for many other areas of social life.⁵ Public officials routinely lie and expect to be lied to. They are not naive subjects in the way that some others might be. This is not to say that public officials deserve to be deceived. Instead, it suggests that, when considering ethical obligations to subjects,⁶ we might be marginally less concerned about lying to politicians than about lying to other subjects *especially* when deceiving them in a minor way uncovers and prevents greater societal harm. Where the outcome of interest is subjects' public duties and where a deceptive experiment is the only way to accurately uncover whether those public duties are being executed in an equitable way, deception may be a small price to pay to defeat discrimination (Riach and Rich, 2004).

That there are compelling reasons to waive informed consent and to use deception with public official subjects, however, does not answer the questions of whether such subjects should be compensated or debriefed. When subjects are ordinary citizens and when no deception is involved, researchers rarely, if ever, compensate subjects for having participated unawares in a field experiment; they also rarely if ever debrief them. Compensation is a tricky issue in the IRB guidelines. When compensation is offered to subjects in advance of the experiment, IRBs are not supposed to take it into account when conducting their risk-benefit analyses of experimental protocols.⁷ The concern is that remuneration can jeopardize consent, prompting potential subjects to agree to risks they would not otherwise undertake. But the IRB guidelines offer nothing specific about whether remuneration can or should be used to compensate subjects for time and resources spent when they did not know they were

Political scientists are, moreover, subject to IRB review, which, though imperfect, allows for a more careful weighing of risks and benefits than journalists or international organizations are subject to.

⁵This argument has been used to justify field experiments on discrimination in the labor and housing markets. The argument is that landlords and employers are not "naive" subjects because deception is a routine component of many ordinary transactions in these markets (Riach and Rich, 2004; Bok, 1978).

⁶I return to other ethical obligations below.

⁷"Direct payments or other forms of remuneration offered to potential subjects as an incentive or reward for participation should not be considered a 'benefit' to be gained from research" (IRB Guidebook, Chapter III, Section A).

part of a study. In the case of field experiments with ordinary subjects, the argument is usually that—because the interventions occur in real-world settings and are designed to be as natural as possible—subjects would be engaging in similar activities on a regular basis in the absence of the research and therefore do not merit compensation. They would be choosing, for instance, whether or not to vote based on similar calculations as those prompted in the experiment. It seems odd to treat public officials—a less vulnerable population—differently from ordinary citizens, especially if, as in field experiments with ordinary citizens, we are observing the conduct of their routine activities.

Some have argued, however, that all types of subjects should be compensated for their participation in experiments of which they were not informed in advance.⁸ Since the IRB regulations are equivocal in this area, the discipline as a whole might try to develop consistent standards one way or the other. If, contrary to current practice, we decide as a discipline that field experiment subjects (both ordinary citizens and public officials) should be compensated, we should bear in mind that public officials do not use their *personal* time and resources during these experiments. Because we are affecting and observing their official duties, public resources are at stake. Following this logic, national treasuries or other public funds would have to be replenished for the time and resources used during the course of the experiment. One might argue for giving money to charities or other non-governmental organizations, but this would constitute a less direct form of remuneration since public officials use public coffers for their daily activities.

Complex and unsettled issues also arise when deception is used with public officials. IRB Guidelines suggest that subjects who have been deceived should be debriefed. At the same time, while the Guidelines counsel that debriefing should be used when it clearly “contributes to the subject’s welfare (i.e., when it corrects painful or stressful misperceptions, or when it reduces pain, stress, or anxiety concerning the subject’s performance)”, they also emphasize that there is greater uncertainty over whether it is appropriate to debrief subjects when such a debriefing could *itself* produce pain, stress, or anxiety..”(IRB Guidelines, Chapter III, Section B; emphasis added). As I discuss further in the next section, there is reason to think that public official subjects might experience negative emotions (annoyance, anger, embarrassment) when told, for instance, that they unknowingly participated in an experiment and were contacted by a fictitious constituent. Debriefing guidelines apply more directly to a laboratory, classroom or doctor’s office, to situations in which subjects knew they were part of research and might have experienced anxiety about their performance during a study because they did not know they were being deceived. In these settings, debriefing can be used as an educational tool, to improve understanding about the value of the research and to improve the welfare of the subject. Field experiments are trickier. They are usually designed to keep subjects in a natural environment, to pose minimal risks to them, and to keep them blissfully unaware of their status as subjects. Informing them of the deception is then likely “itself to produce pain, stress or anxiety.”

Indeed, to my knowledge, none of the field experiments that have been conducted to date with public official subjects have debriefed the subjects afterwards. My original application for IRB review of the experiment I conducted with South African politicians—which involved deception—included a

⁸See, for instance, Andrew Gelman’s blog entry on May 6, 2010 and the ensuing discussion: http://andrewgelman.com/2010/05/63000_worth_of/.

proposal for a full debriefing that would have explained to the subjects that they were part of the experiment, explained the nature of the experiment and its justifications and privacy procedures, stated initial findings and given the subjects contact information for me and my institution's IRB. Before granting approval, however, my IRB asked that I remove the debriefing portion of my protocol. I have had some back and forth with the IRB about its reasoning, and the discussion is ongoing. Because the IRB Guidelines are equivocal about debriefing in some contexts, there is clearly a grey area here. The IRB is not wrong, according to the guidelines, to lean towards not debriefing subjects. Rather since the Guidelines are not clear, we need continued debate until we can agree on more precise standards. Subjects in my experiment and others like it were observed in routine behavior. Debriefing might indeed have produced negative emotions that would not have otherwise arisen. Public officials are also not students, and debriefing might not have accomplished the educational goals that the guidelines intend it to accomplish. (But see the discussion below.) At the same time, in my view, fully respecting subjects of any type means weighing the wrong done to them through deception more heavily than the pain that they might experience knowing the truth.

Responsibilities to Researchers

Thus far I have considered our responsibilities only to the subjects themselves. Yet it is important to remember that public officials constitute a more limited subject pool than ordinary citizens. Unlike ordinary citizens, public officials also control funding for many universities and for much research. Many political scientists would like to study public officials and need to use public funding to do so. What then are our ethical obligations to *other researchers* when we conduct field experiments with public officials? Do some types of experiments risk jeopardizing funding for future studies? If deception or the waiving of informed consent upset subjects, will such design elements make politicians less likely to cooperate with and support future political science research? Should there be a stronger prohibition on these design elements when using public officials as subjects?

We might ask first whether we know anything about whether the waiving of informed consent and deception upset subjects or alter their willingness to participate in future research. To my knowledge, there have been no studies of how public officials respond to the use of deception or the waiving of informed consent. But, based on studies with undergraduate students and other citizen subjects, there is some cause to be cautious. Some psychology studies, for instance, have suggested that subjects who learn they have been deceived or studied unawares experience negative emotions, such as embarrassment, annoyance, betrayal and anxiety (Fisher and Fyrberg, 1994; Finney, 1987; c.f. Christensen, 1988; see also Ortmann and Hertwig, 2002 for a summary). Finney (1987) even found that undergraduate subjects who were deceived saw less value in the research. Academics who have been the unknowing subjects of deceptive research have also expressed annoyance and even outrage.⁹ If public officials discover or are debriefed about our field experiments and respond similarly, there is some chance they might perceive political science research more negatively than they did before.

Would they retaliate or be less willing to participate in future research? Again, our evidence is weak but calls for some caution. Jamison, Karlan and Schechter's 2008 study is one of the few that has

⁹See, for instance, Andrew Gelman's blog entry on May 6, 2010 and the ensuing discussion: http://andrewgelman.com/2010/05/63000_worth_of/.

looked directly at the impact of deception on subjects' willingness to participate in future research. The authors had subjects play an ultimatum game. Subjects in the treatment group were led to believe they were playing against real people when in fact they were playing against computers. A control group was not deceived. Having debriefed the deceived subjects, the authors tried three weeks later to recruit all subjects into an unrelated experiment. While they found few significant differences in turnout between the two groups, they did find that female subjects who were deceived were slightly less likely to return than non-deceived female subjects. Subjects who had received less money in the first experiment were also somewhat less likely to return if they were deceived. Although without further research we cannot be sure, these results suggest that there is some chance that deception impacts subjects' willingness to participate in future research. Taking this chance into account, should we prohibit deception and the waiving of consent with public official subjects?

On the one hand, to place prohibitions on research with powerful subjects just because they are powerful seems wrong. The behavior of public officials has a wide impact on many individuals and groups. For this reason we should want accurate and rich information about the drivers of such behavior and about whether such behavior is accountable and equitable. We would be failing as a field if we had to restrict certain research methods to the weak. On the other hand, for the same reason, we should want such research to continue far into the future, and we should recognize that public funding and continued access is critical to such research. The task then is to figure out ways to minimize the externalities.

One rule of thumb is that researchers should conduct field experiments in countries and with public officials whom they expect and need to study again. Doing so internalizes some of the potential costs. Scholars are more likely to conduct deceptive field experiments and to waive informed consent only when absolutely necessary if they risk their *own* future funding and research. This rule of thumb is not easily enforced, but it is an ethical standard that we might use to judge our colleagues and ourselves.

At the same time, we should think carefully about debriefing methods and public advocacy efforts. If there is great societal value to research that, under the right circumstances, uses deception and waives informed consent with public official subjects, then we need to make that value clear to the powers that be. This might mean that, if we decide to debrief, we develop debriefing methods that are truly educational in nature, rather than impersonal and one-shot. For instance, debriefing after field experiments often means sending individual emails to subjects informing them of the nature of the research. This method often does not allow a dialogue that communicates the value and ethical considerations of the experiments. Researchers might instead invite subjects to an online forum with both a video of the researchers explaining (and justifying) the research in detail and space for subjects to chat with the researcher, ask questions, express their reactions and receive responses anonymously. The idea would be to create a debriefing environment that is as similar to a classroom as possible.

Making clear the social value of our field experiments should also mean developing a better discipline-wide advocacy effort that regularly engages with politicians in sharing and justifying our methods and findings. Given that members of the US Congress have recently (and repeatedly) introduced bills to cut political science funding, this type of advocacy is probably necessary whether we are talking about field experiments with public officials or not.

While field experiments that use public official subjects are important because they scrutinize important political behavior in invaluable ways, the ethical issues raised by such experiments are complex. We need ongoing debate about whether (and how) such subjects should be compensated and debriefed. We also need ongoing debate about effective ways to minimize externalities imposed on other researchers by such experiments. I hope this notes has made some progress towards beginning those conversations.

References

- Bertrand, M., and S. Mullainathan. 2004. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94 (4): 991-1013.
- Bok, S. 1978. *Lying: Moral Choice in Public and Private Life*. New York, NY: Pantheon Books.
- Brown, C. and P. Gay. 1985. *Racial Discrimination 17 Years After the Act*. London: Policy Studies Institute.
- Butler, D. and D. Broockman. 2011. Do Politicians Racially Discriminate against Constituents? A Field Experiment on State Legislators. *American Journal of Political Science* 55 (3): 463-477.
- Christensen, L. 1988. Deception in Psychological Research: When is its Use Justified? *Personality and Social Psychology Bulletin* 14: 664-675.
- Finney, P. 1987. When Consent Information Refers to Risk and Deception: Implications for Social Research. *Journal of Social Behavior and Personality* 2: 37-48.
- Fisher, C. and D. Fryberg. 1994. Participant Partners: College Students Weigh the Costs and Benefits of Deceptive Research. *American Psychologist* 49: 417-427.
- Humphreys, M. 2011. Ethical Issues for Embedded Experimentation. *Comparative Democratization Newsletter*, October 9(3): 10.
- Humphreys, M. and J. Weinstein. 2012. Policing Politicians: Citizen Empowerment and Political Accountability in Uganda Preliminary Analysis. Working Paper.
- Jamison, J., D. Karlan, and L. Schechter. 2008. To Deceive or Not to Deceive: The Effect of Deception on Behavior in Future Laboratory Experiments. *Journal of Economic Behavior and Organization* 68(3): 477-488.
- Loewen, P. and M.K. MacKenzie. 2011. Representation in a Federal System: a Field Experiment. Working Paper.
- Loewen, P. and D. Rubenson. 2011. For Want of a Nail: Negative Persuasion in a Party Leadership Campaign. *Party Politics* 17(1) 45-65.
- Loewen, P., D. Rubenson and L. Wantchekon. 2010. Help Me Help You: Conducting Field Experiments with Political Elites. *The Annals of the American Academy of Political and Social Science* 628: 165-175.
- Malesky, E., P. Schuler, and A. Tran. 2011. The Adverse Effects of Sunshine: Evidence from a Field Experiment on Legislative Transparency in an Authoritarian Assembly. Working Paper.
- McClendon, G. Co-ethnicity and Democratic Governance: A Field Experiment with South African Politicians. Working Paper.

- Ortmann, A. and R. Hertwig. 2002. The Costs of Deception: Evidence from Psychology. *Experimental Economics* 5: 111-131.
- Pager, D., B. Western and B. Bonikowski. 2009. Discrimination in a Low Wage Labor Market: A Field Experiment. *American Sociological Review* 74: 777-799.
- Riach, P. and J. Rich. 2004. Deceptive Field Experiments of Discrimination: Are they Ethical? *Kyklos* 57(3): 457-470.
- Roth, A. 2001. Form and Function in Experimental Design. *Behavioral and Brain Sciences* 24: 427-428.
- Wantchekon, L. 2003. Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin. *World Politics* 55: 399-422.

Ethics and Politics in Field Experiments

Kyle Peyton

Research Support Officer at the Melbourne Institute of Applied Economic and Social Research
University of Melbourne
kyle.peyton@unimelb.edu.au¹⁰

The country needs and unless I mistake its temper, the country demands bold, persistent experimentation. It is common sense to take a method and try it: If it fails, admit it frankly and try another. But above all, try something

-Franklin D. Roosevelt, 1932

What is needed, what is demanded, and what is done are important issues for experimentalists. One might be easily convinced that classical experiments are the 'gold standard' for evaluating program effectiveness and that we need more of them. The classic texts make convincing methodological arguments on these grounds in the first chapters (see, Riecken and Boruch, 1974; Orr, 1999; Shadish et al., 2002). Yet, in the context of field experiments that require government and community support, demand and implementation may be more important issues and politics and ethics the most salient concerns.

Heckman and Smith (1995) suggest that ethical and public relations concerns are the primary barriers. According to a survey from Doolittle and Traeger (1990) 'random assignment' and 'denial of services to controls' were cited by 61.8% and 54.4% of service delivery areas (SDAs) contacted about possible participation in the National Job Training Partnership Act (JTPA) Study. Although not particularly encouraging, these objections may have been driven by an inadequate understanding of

¹⁰I'm grateful to the participants at the 2012 Methods of Evaluating Programme Impacts course for candid responses to my survey questionnaire and to Andrew Leigh and John Micklewright for their conversations and advice. Thanks also to Eric Coleman, Laura Langbein, Greg McAvoy, Rebecca Morton, Cyrus Sammii, Dawn Teele and Dustin Tingley for responding to the question I sent to the Political Methodology list in February 2012 which was the catalyst for this article. All data used will be made available on request.

ethical issues. One criticism of social scientists has been insufficient ethics training and an inability or unwillingness to engage in public discourse on the topic (Blustein, 2005).

Despite a boom in experimental methods across the social sciences, there has not been a corresponding increase in experiments involving government partnership. Indeed, 'ethical concerns' are sometimes viewed as a euphemism for political and ideological objections. A cynical view is that opposition to experiments is driven by fear of evaluation rather than a genuine concern for ethics (Palmer and Petrosino, 2003). But is it really the case that ethical 'concerns' are the primary barriers to experimentation, as the JTPA results suggest, or is it more complicated?

This article briefly examines ethics and politics in the field of experimental political science. Despite finding a steady output of experimental articles over the last six years, there is almost no reference to ethical issues and only one example of an experiment with government partnership. To shed light on potential concerns, results from a small survey conducted of public servants and applied researchers are presented. The responses provide no support for the conventional wisdom that randomization and withholding treatment are the primary concerns or that political barriers are insurmountable. The article then reviews literature on ethics with the Belmont Report as a guide. Finally, politics and engagement in the field are discussed. Although practical constraints are barriers to 'ideal experiments' there is no shortage of guidance for those who want to *try something*.

Taking stock

It's fair to say that 'experimental political science' has grown tremendously in popularity.¹¹ Druckman et al. (2006) and Morton and Williams (2008; 2010) have documented and discussed the emergence, growth and historical trends of experimental articles with Morton and Williams (2010) outlining many alternatives to the ideal experiment and Druckman et al. (2006) providing evidence that experimental articles are more likely to be cited. Continuing in this tradition, I examined the frequency of experimental articles by type in three leading political science journals, *The American Political Science Review*, *The American Journal of Political Science* and the *The Journal of Politics* over the time period 1 January 2006 to 31 December 2011.¹²

¹¹see Greenberg et al. (2003) for a history of social experimentation and its influences on policy

¹²I recognize this is a biased sample in the sense that it underestimates the number of field experiments in the discipline. I'm interested in the relative balance of field experiments between alternatives and I think this is an OK sample for that purpose.

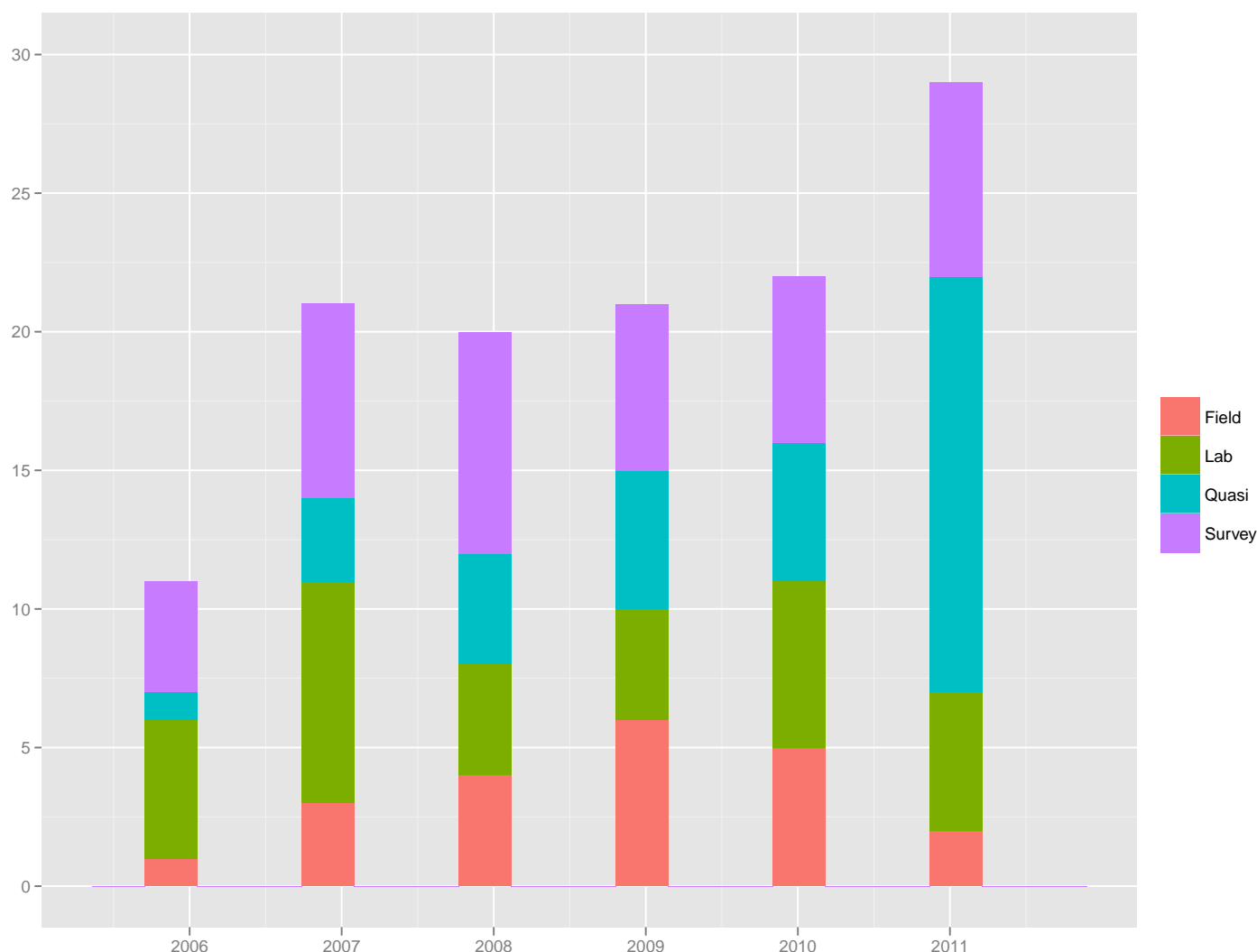


Figure 1: Experimental articles by type. *Notes:* I conducted a separate search in each journal using EBSCOhost. The initial search was restricted to articles that contained the word ‘experiment’ or variants such as ‘experimental’ or ‘experiments’. The initial search identified 133 articles; of those, 124 were identified as relevant and classified into four groups following the definitions used in Morton and Williams (2010). Estimates are, of course, downward biased and only used as anecdote.

Figure 1 shows a five year trend in experimental articles. The upturn of quasi-experimental¹³ articles from 9% of the total in 2006 to 52% of the total in 2011 is worth noting. The data also show a decline in field experiments since 2009. This relative ‘underprovision’ provides support for Green and

¹³I define ‘quasi-experiments’ as those where the researcher does not interfere in the data generation process but instead uses experimental techniques such a regression discontinuity design, matching methods, ‘difference-in-differences’ or instrumental variables in natural experiments

Gerber’s (2003) critique. Of the 124 articles returned for the search period only 17% involved field experiments whereas 26% involved lab experiments, 30% used survey experiments and 27% relied on quasi-experimental methods. Of the articles on field experiments, three were secondary analyses and therefore did not involve manipulation of the data generation process (DGP) by the researcher. Only 1 of 18 *new* field experiments involved a government partnership, Benjamin Olken’s (2010) direct democracy experiment in Indonesia. Absent from almost all of the articles was any discussion of ethics.

Figure 2 shows a word count of some common words appearing in the articles sampled over the period. Although it is reasonable to expect that none of the quasi-experimental articles need pass scrutiny of an Institutional Review Board (IRB), this is not necessarily true for survey, lab and field experiments which generally involve manipulation of the DGP by the researcher. Yet, despite widespread use of experimental words such as ‘treatment’, ‘control’ and ‘random’ since 2006, only four of the articles—three field experiments and one lab experiment—discussed ethical issues associated with the experiment conducted.

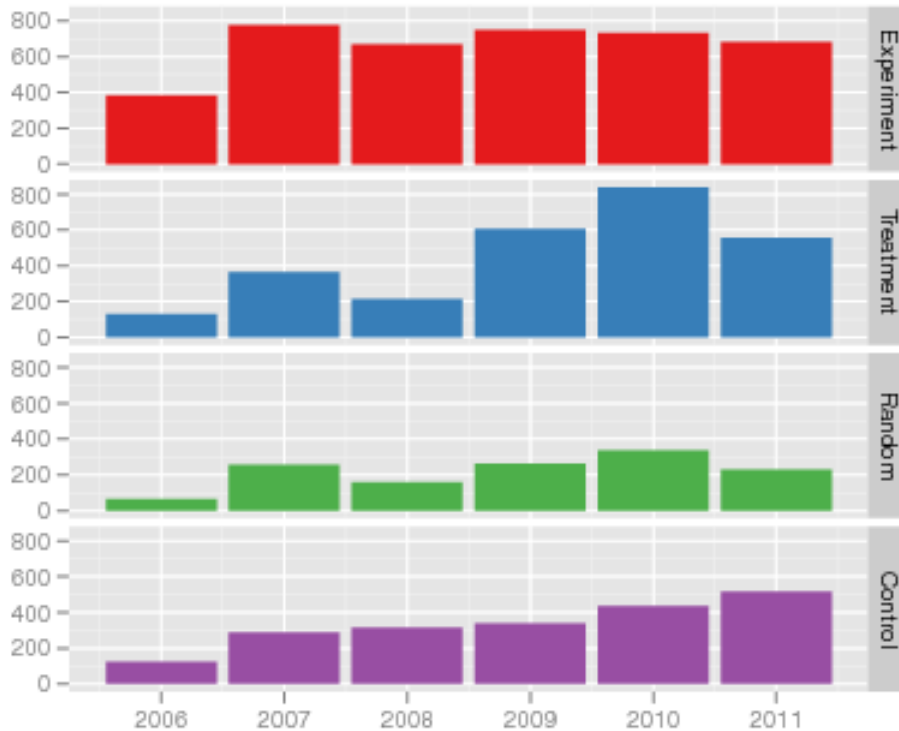


Figure 2: Words in experimental articles. *Notes:* I examined all 123 articles using a free program called ‘Word Counter’ available for Macintosh operating systems.

One notable field experiment testing the effects of mass media on political culture in Rwanda provided an appendix devoted to ‘procedural and ethical protocol’ (Paluck and Green, 2009). Another, involving an experiment on public officials contained a section on ethics in the main text (Butler and

Broockman, 2011). Gerber et al. (2008) explain in a footnote that the research was reviewed and approved by an ethics committee and Boudreau and McCubbins (2010) use a footnote to explain how a committee ruled it was unethical to force subjects to pay money for polling information in a lab experiment.

Not all experiments are created equal. The concentration of ethical conversation supports the argument that field experiments are generally the most ethically questionable (Teele, *forthcoming*). Because the field setting involves direct interference in the DGP, experiments may require the most justification from the research ethics body and the community for several reasons. The remainder of the article discusses many of these reasons. An important first question for researchers is, what might these objections be?

Anecdote on Attitudes

Aside from Heckman and Smith’s 1995 paper, little evidence exists to prime researchers for objections they might encounter. Leigh (2003) outlines several concerns raised toward ‘randomized policy trials’ that are likely to surface in the field. Among these are the traditional concerns about randomization and withholding treatment reported in Heckman and Smith as well as others such as “[experiments] don’t work because it is too difficult to define the goals of most policies”. Table 1 provides responses from a survey conducted in March 2012 on some commonly raised ethical concerns:

Question	Agree %	Disagree %	Don’t know %
Random assignment is unethical	12	88	0
Withholding treatment is unethical	20	76	4
Experiments are unethical without informed consent	64	28	8
Experiments are ethical if benefits outweigh costs	44	48	8

Table 1: Traditional ethical concerns. *Notes:* Survey of applied researchers and public servants at the 2012 Melbourne Institute *Methods of Evaluating Programme Impacts* course in Canberra, Australia from 26-27 March 2012. Sample includes public servants and applied researchers from wide range of institutions and seniority levels. Four respondents worked at universities and the remainder worked in government organizations. N=25/27.

Among those surveyed, randomization and withholding treatment were not the primary ethical concerns. In fact, only three respondents agreed with the statement “randomized controlled trials are unethical because they involve random allocation of services”. This was the only question where no respondents ticked ‘I don’t know’, suggesting views about randomization were well defined. The majority of respondents also disagreed that experiments were unethical because they involved withholding treatment from controls.

On the other hand, most agreed that experiments were unethical without informed consent. Respondents were divided on the issue of whether or not experiments are ethical if the benefits

outweigh the costs. According to Morton and Williams (2010, p. 455) this is one of the most important ethical considerations for IRBs. Table 2 shows responses to questions related to general attitudes and the institutional environment respondents were operating in.

Question	Agree	Disagree	Don't know
	%	%	%
Experiments most effective	12	64	24
Objections primarily political	36	40	24
Procedural guidelines clear	8	68	24
Ethical guidelines clear	20	52	28
Political concerns can be alleviated	64	20	16
Ethical concerns can be alleviated	84	8	8

Table 2: Attitudes toward experiments. *Notes:* Survey of applied researchers and public servants at the 2012 Melbourne Institute *Methods of Evaluating Programme Impacts* course in Canberra, Australia from 26-27 March 2012. Sample includes public servants and applied researchers from wide range of institutions and seniority levels. Four respondents worked at universities and the remainder worked in government organizations. N=25/27.

Although the majority agreed that ethical and political concerns can be alleviated only three believed classical experiments were the most effective research design. This suggests researchers may encounter objections to the ideal experiment even if objections to random assignment and withholding treatment are not raised. Moreover, despite optimism toward ethical and political barriers, only a small minority agreed that organizational guidelines on ethics and procedure were clear. This suggests researchers can benefit from proactive engagement with partners, especially when projects require approval from IRBs outside the university.

Although designed primarily for biomedical research, the document used by most IRBs in the United States is the Belmont Report or some variation. The Report concentrates on three principles fundamental to ethical research with human subjects: *Respect*, *Beneficence*, and *Justice*. These principles also have international relevance. The National Statement on Ethical Conduct in Human Research used in Australia, for example, concentrates on the same basic principles outlined.¹⁴ The next section explores these basic ethical principles and then surveys academic literature related to randomization and withholding treatment that may assist those who engage more directly with their research partners.

Informed Consent, Deception and Basic Ethical Principles

In a Kantian sense, subjects should not be treated as objects; that is, researchers should not observe subjects unless subjects authorize the observation.

-Corsi, 1983

¹⁴See Morton and Williams pp. 419-421 for discussion of research ethics bodies in other countries

One of the most common ethical concerns in experimental research is ‘informed consent’.¹⁵ Indeed, the survey conducted suggests this may be most important ethical issue with 64% of respondents agreeing that randomized controlled trials are unethical without informed consent. This concern is also part of the Belmont Report’s first ethical principle, *Respect for Persons*.¹⁶

This most basic interpretation requires that subjects are treated as autonomous agents and those with diminished autonomy are entitled to protection. According to the Report, “an autonomous person is an individual capable of deliberation about personal goals and of acting under the direction of such deliberation. To respect autonomy is to give weight to autonomous persons’ considered opinions and choices while refraining from obstructing their actions unless they are clearly detrimental to others.” Moreover, “the selection of research subjects needs to be scrutinized in order to determine whether some classes are being systematically selected simply because of their easy availability, their compromised position, or their manipulability”. This principle raises some ambiguity and is a well documented problem in the history of social research.

Morton and Williams (2010) provide a sobering reminder of this history since the 1947 Nuremberg Code. They also provide guidelines for establishing informed consent and devote a chapter to the ethical issues associated with deception.¹⁷ A visceral example that captures the essence of the *Respect for Persons* principle occurred at the infamous Willowbrook State School in New York. Through a series of experiments, Krugman et al. (1967) fed stool extract to mentally handicapped children to study hepatitis. The ethical merits of these experiments were debated at length in several 1971 issues of *The Lancet* known as the ‘Willowbrook Letters’.¹⁸ One of the issues that came up in the Letters was whether or not a code of ethics should be interpreted as providing guiding principles or as a “legal document to be searched for convenient loopholes” (Shapiro, 1971).

There is good reason to believe researchers should be concerned with both. As a code of ethics, the Report provides researchers with flexibility. On the topic of informed consent, for example, the Report states, “a simple listing of items does not answer the question of what the standard should be for judging how much and what sort of information should be provided.”¹⁹ The Report also acknowledges pragmatism and recognizes situations in which subject knowledge may impact the validity of research. Indeed, so called ‘Hawthorne Effects’ criticisms have been used as justifications for deception in the past (see Banerjee and Duflo, 2009; Levitt and List, 2009).

The Report acknowledges that “in many cases it is sufficient to indicate to subjects that they are being invited to participate in research of which some features will not be revealed until the research is concluded.” The Report also provides guidelines for ‘comprehension’ and ‘voluntary consent’. Here, the guidance closely mirrors conditions under which a contract is deemed *voidable* in common law

¹⁵See Miké 1989 for a discussion of the Categorical Imperative and its application to randomized clinical trials and complex statistical methodology.

¹⁶<http://ohsr.od.nih.gov/guidelines/belmont.html>

¹⁷For discussion of ‘embedded experiments’ see Humphreys (2011) and Teele (*forthcoming*)

¹⁸The interested reader can consult Beecher (1971), Edsall (1971), Goldby (1971), Krugman (1971), Pappworth (1971), Shapiro (1971) for the full debate.

¹⁹ see <http://ohsr.od.nih.gov/guidelines/belmont.html>

countries.²⁰ As Michael Oakes explains, “an informed consent document is a binding legal contract that permits remedy through not only tort but also contract law” (2002).

The definition of ‘informed’ is crucial.²¹ Lidz and Roth (1983) provide guidance to the legal and ethical complexities and explain why signing a form is a necessary but far from sufficient condition for establishing informed consent. Corsi (1983) provides examples of legal and ethical issues that surface when the experiment requires support from many stakeholders, as most will. Even if participants decide informed consent has been violated *after* the study problems can arise. For example, the Australian Longitudinal Study of Health and Relationships was embroiled in scandal after a participant complained on public radio and took legal action against the researchers.²²

Respect for Persons, however, is not the only concern. The principles of *Beneficence* and *Justice* are also worth consideration. Morton and Williams (2010) provide a modern treatment of these principles in the language of ‘Institutional Review Board (IRB) speak’ with acknowledgement of the complexities and exceptions the general formula does not cover. Unique to their discussion is the concept of ‘minimal risk’ as a potential vehicle for expedited approval. *Beneficence* is defined by two guiding rules: 1) do no harm, regardless of potential benefits and 2) maximize *possible* benefits and minimize *possible* harm.

Of course, any code of ethics is open to some interpretation. Because researchers need to make value judgements without the ability of forecasting the future thinking about risk is necessary. Specifically, what is the *probability* and *magnitude* of benefit and harm? The airplane example provides a useful illustration. Consider a flight from Chicago to New York. The probability of a crash on any given day is approximately zero; however, if the plane does crash the result is death for all passengers with near certainty. Here, the magnitude of harm is large although the probability of realizing that harm is approximately zero.

To make things even more equivocal, the probability of ‘harm’ is 1 in any experiment depending on how the term is defined. For example, the opportunity cost associated with participation is always present and is in fact one justification for paying research subjects to participate. To address this ambiguity, Morton and Williams discuss three ‘standards for comparison’ to consider when minimizing risk: 1) daily life, 2) performance of routine physical examinations and 3) performance of routine psychological examinations.

Unlike medical experiments and the airplane example, the probability of death or serious physical harm in most social science experiments is approximately zero. For field experiments, the first standard for comparison is most relevant. The risk of harm is typically about the same as that of daily life. As Morton and Williams explain, “a field experiment would provide a subject with minimal risk if the

²⁰The principles in contract law are duress, undue influence and unconscionable bargaining. The first involves either physical threats or unacceptable commercial pressure so as to procure agreement. The second involves the use of a position of power or influence over another. The third involves one party taking advantage of another, possibly due to that party’s intellectual vulnerability or lack of understanding. See: <http://law.anu.edu.au/colin/Lectures/uncons.htm>

²¹Special issues may arise in developing countries without a common law legal culture. For examples and suggestions see Morton and Williams p.493-5.

²²For more details see a full transcript of the participants interview by the Australian Broadcasting Corporation here: <http://www.abc.net.au/radionational/programs/healthreport/research-ethics/3320450>

experiment had an equivalent probability and magnitude of harm of the subject as events in his or her daily life.” Yet not all harms are created equal.

The most difficult to justify and the least relevant are physical harms. Most relevant for social science field experiments are psychological harms and social and economic harms. Invasion of privacy and violation of confidentiality, although historically significant, are much less relevant today given the advent of modern technology.²³ Morton and Williams (p. 410) also provide a useful discussion of *Justice*, which is closely related to potential benefits: societal benefits, therapeutic benefits and collateral benefits.

As they explain, *Justice* is fundamentally about the distribution of these benefits. Researchers should think about who receives the benefits of research *and* who bears the costs. Orr (1999) has even proposed a mathematical heuristic:

$$EV = \pi * V_p - C$$

EV is the *expected* value of the experiment. V_p , the value of a change in policy, is the net social benefit that will accrue if the experimental program is adopted, π is the probability of change in policy and C is the cost of the experiment. Orr argues that experiments have no social value unless their results influence policy, but fails to consider how these benefits should be distributed.

According to the Belmont Report, ‘distribution’ can take on several interpretations. The choice of interpretation is likely driven by context. In medicine, for example, potential benefits may be allocated on the basis of individual need rather than some measure of effort. Social scientists, on the other hand, might be more interested in the equal share principle. Again, the Report serves as a guide rather than a rigid regulatory apparatus.

Obviously this discussion does not capture the gamut of ethical and legal issues related to experiments nor does it attempt to. Instead, it serves as a primer on some key issues. The finer details are borne out in application to particular circumstances. Margetts and Stoker (2010) provide a quick and accessible introduction to experiments with a section on ethics and a focus on political science examples. Morton and Williams (2010) gives a more detailed treatment and contains three chapters on ethics.

For the forward looking researcher, two useful texts on ethical and legal issues are also worth consulting: *Ethical Issues in Social Science Research*, Beauchamp et al. (1982) and *Solutions to Ethical and Legal Problems in Social Research*, Boruch and Cecil (1983). Both texts provide a rich source of *potential* issues and strategies to overcome them. Most of the contributors speak with the benefit of hindsight and years of experience. Much can be learned from exploring the mistakes others made and reading their candid confessions. The next section concentrates on two well documented implementation concerns: randomization and withholding treatment.

Random Assignment and Withholding Treatment

²³For the interested reader, Gary King (<http://gking.harvard.edu/>) and Latanya Sweeney (<http://dataprivacylab.org/people/sweeney/>) both have webpages devoted to informatics and data sharing. Thanks to Molly Roberts for providing the link to Latanya Sweeney.

How can a computer tell me whether or not I need treatment?

-Anonymous participant from Lam et al. (1994)

In the context of experiments, the word ‘random’ is a pejorative. In colloquial usage it means *haphazard*. Regardless of whether or not the benefits are potentially beneficial separation raises questions about fairness. Hence, withholding treatment is a common source of objection to the classic randomized controlled trial.

This has been documented and discussed since at least Campbell (1974). This *will* impact research design and may become a crucial issue for feasibility review (See Farrington and Jolliffe, 2002 and Campbell, 2003). The ‘man (or woman) on the street’ may not understand that random assignment means all individuals have an *equal* probability of selection. As King et al. (2007) explain “researchers often think of randomness as the ultimate in fairness, because the assignments it produces would be the same even if the name of the person receiving the random number changed; but this anonymity property is often viewed by citizens and politicians as the ultimate in whimsy, ignoring as it does how much the person not receiving the services needs them.”

Most academic discussions, however, place the benefits of experimental design over ethical concerns. Burtless and Orr (1986), a classic experiments reference, argue that objections to experiments on ethical grounds are simply not *valid*. Leigh (2003) argues that ethical objections to random assignment fail to consider “the fact that governments never provide assistance to all those who would benefit from it”. When resources are scarce, argue Banerjee and Duflo (2009), randomization is the fairest rationing mechanism.²⁴

In the jargon of economics, there is an ‘asymmetry problem’. Even though researchers believe this, that does not mean everyone in government and the community feel the same way. Although criticism is often directed at ‘bureaucrats’, the survey conducted suggests not all those involved in policy implementation have the same views. Indeed, most agreed that ethical concerns can be alleviated. Yet this does not mean addressing concerns is unnecessary. Failing to do so may mean failure of the project. Ideally, design issues will be explained and agreed to as part of the informed consent process.

Researchers can “learn how to respond to ethical concerns” (Lowen et al., 2010). Once it has been explained that randomization means everyone has an equal probability of being selected it seems much more reasonable. One recommendation from a senior policy advisor who completed the survey was that academics should stop using the word ‘randomization’ altogether as it is completely misunderstood. A more pragmatic suggestion is that ‘randomization is always acceptable at one level below that at which politicians care’ (King et al., 2007).

Assuming researchers do randomize, however, how should it be done? A coin flip? Drawing a name from a hat? A giant Lottery Ball spinner?²⁵ One thing seems clear: avoid as much human error as

²⁴If the alternative is patronage or nepotism this seems like a reasonable argument. For an alternative perspective see Teele (*forthcoming*)

²⁵In the documentary *Waiting for Superman* allocating school vouchers to an eager audience of poor children on the basis of a giant ball spinner certainly made the whole scene more unpleasant than it needed to be. The additional psychological harm caused by this exercise seemed completely gratuitous.

possible. Arguments for computerized randomization are often made on these grounds alone. Although it reduces the temptation to make exceptions for treatment and the possibility of human error one might also argue that the computerized process is more ethical because it reduces stress on the researchers and other members of the experimental team. In the same way that pressing a lever is less stressful on the executioner than bringing down an axe by hand, having a computer ‘decide’ whether or not an individual gets treatment on the basis of some pseudo-random²⁶ process may reduce harm for administrators of the treatment. The end is the same but technology provides a more palatable means.

A pure interpretation of the Belmont Report demands that considerations of potential harm extend to *all* participants, not just the research subjects. Morton and Williams have also recommended that researchers consider harm to non-subjects. That is, the harm to researchers and ‘confederates’ as well as harm to third parties or societal risks more broadly.²⁷

Satisfied confederates are especially important. This moral stress is both a potential psychological harm and has been shown to directly impact the efficacy of the research project. Dee and Keys (2004), for example, discuss problems related to teachers complaining that separation reduces morale in a pay for performance experiment. Likewise, administrators involved in giving the treatment who are opposed to the idea of randomization at the onset they may deliberately circumvent the process even if the importance has been explained to them (Torp et al., 1993).

It is impossible to separate the ethics and politics of field experiments. The forward looking researcher can benefit from an understanding of how these concepts operate in tandem. In a special issue on ‘Randomization and field Experimentation’ in the Winter 1985 issue of *New Directions for Program Evaluation* Boruch and Wothke (1985) argue, “it is not easy to design the randomization process so that it is robust against indifference, incompetence, and corruption. Still, it is a task that seems worth the effort.” The next section outlines some political problems that may arise during the experiment.

Politics and engagement

The debate shouldn’t just be academics trying to convince ‘stupid/ignorant’ public servants why RCTs are good – we already know.

- Anonymous survey respondent

A social experiment is a political act. Although political factors have been discussed at length since at least Riecken and Boruch (1974), public acceptance is less understood. Hillis and Wortman (1976) found that attitudes toward experiments can be substantially improved depending on how the research project is described. Lam et al. (1994) provide tangible evidence of ‘framing effects’ across different stakeholders related to a drug rehabilitation project in New Haven, CT. They caution that “very serious attention needs to be paid to explaining and re-explaining the randomization process and the

²⁶Because computer number generators are not purely ‘random’ there may be a really interesting philosophical debate hiding in here.

²⁷For a useful example of ethical issues facing confederates see the discussion of a bribery experiment in Mexico on pp. 471-2.

importance of using an experimental design to all relevant parties.” In other words, engagement is important.

According to Lowen et al. (2010), “experiments should not be viewed as one-shot affairs but instead as the beginning of a relationship”. This requires transparency and engagement with *all* stakeholders from the onset. Field experiments are not easy and political problems are the norm rather than the exception. Even the famous STAR experiment was not the ideal type. It involved ‘re-randomization’ in response to complaints by parents and organized lobbying (Krueger, 1999).

Actions taken by third parties can severely damage the experimental design. For example, in a subsidized meals program for schools in Kenya, more than *half* of the control schools organized funds to implement meal programs of their own (Vermeersch and Kremer, 2004). In another instance, union members blocked an experiment involving the privatization of public goods (Bothm, 1984).

Even though political problems will emerge there is good reason to be optimistic. People have been thinking about these issues for decades, most problems are not new and a great deal of *practical* solutions have been developed to modify the design process. Social scientists have been particularly good at developing ‘scalable’ Nickerson (2005) and ‘politically robust’ King et al. (2007) design solutions to overcome classic political problems.

For the interested reader, a special issue on experimental methods in the *ANNALS of the American Academy of Political and Social Science* contains more discussion on related issues with several examples of real implementation challenges. Cotterill and Richardson (2010) also provide examples of problems that may occur at the local level when government is a partner. Boruch and Wothke’s ‘Seven Kinds of Randomization Plans for Designing Field Experiments’ is a great primer and provides deep insight and wisdom. Two timeless and accessible sources of information on design issues are Orr (1999) and Shadish et al. (2002). Strategies have even been outlined for developing written agreements with partners Fairweather and Tornatzky (1977). And should researchers need reminding of what’s possible, King et al. (2007) describe the largest field experiment in history and provide guidance for those navigating the politics of experimental design.²⁸

Conclusions

My aim in suggesting an ethics of evidence has not been to propose yet another code, when codes have so often been notoriously ignored. Change occurs when conditions are right and there is widespread awareness. My aim is to speak of an ethics of evidence as a means of consciousness raising, of increasing awareness. There is a strong ethical basis for developing the best possible evidence for medical procedures, and the time may be right for a wider hearing.
- Valerie Miké (1990)

The same may be said for social research. One side effect of ‘whispering in the ears of princes’, might be an accumulation of evidence to develop the best possible social policies (Roth, 1995). As *The Economist* has suggested, why not, try it and see? That is, put in place a policy and evaluate it to

²⁸There are also freely available slides and other materials at King’s website: <http://gking.harvard.edu/>.

determine what works (Campbell, 1969). One might ask, do the costs of bad public policy outweigh the ethical issues associated with social experiments?

No doubt, much can be gained from pilot studies as pre-requisites for policy initiatives and evaluating existing programs for effectiveness. These were among the chief justifications for experiments in Orr (1999). Challenging conventional wisdom is always a good idea. The harmful consequences of the 'Scared Straight' program revealed by randomized controlled trials serves as one notable example (Petrosino et al., 2003). Many such assessments of 'What works?' have been the focus of the Campbell Collaboration's initiative on evidence based policy.²⁹

One potential danger of 'beating the drum' for experiments, however, is that they become a tool used solely for career advancement. This has been documented as an ugly blemish in the history of medical research (Pappworth, 1968). Social scientists have the benefit of hindsight and can avoid the same fate. Professional ethics can serve as a guiding light and reference point for how researchers conduct themselves, especially when the research potentially affects the lives of others, sometimes millions, in a very tangible way.

For example, is an intervention that exogenously modifies class size a good idea because it will lead to publication in a top journal or because additional studies about the effects of class size on learning will likely add more than is already known from Krueger's experiment? If the answer is the first, it may be better to go back and analyse the data again. As Sondheimer and Green (2010) show, data from the same experiment can provide new answers to important research questions.

Ethical violations may also have profoundly damaging consequences. As Lam et al. (1994) make clear, the reputation of an institution or the research team is an important factor in recruiting research subjects. If town and gown relations are strained, the likelihood of conducting a local field experiment is low. Moreover, if a project is suspended for ethical violations, funded research is likely to be suspended. Oakes (2002) provides a sobering warning of the consequences:

Participants cannot receive treatments, enrol, or be recruited; results from time-sensitive studies cannot be reported; and data cannot be analyzed...there is no money to pay graduate students, travel to conferences, or purchase equipment...researchers may lose months, if not years, of work. Severe effects to an institution's reputation may dislodge the public's willingness to participate in research ... ethically improper research, by anyone, devastates a social scientist's chance to recruit from affected communities long into the future.

Field experiments present unique ethical and political challenges. As Teele (*forthcoming*) makes clear, researchers must look beyond the ethical/unethical dichotomy and embrace contextual factors when considering ethical implications of field experiments. This does not mean researchers shouldn't experiment. The Belmont Report serves as a flexible guide to ethical conduct rather than a hindrance. Trade off's between ethics and evidence can be viewed from a holistic perspective with all stakeholders in mind.

²⁹<http://www.campbellcollaboration.org/>

Experimental methods are growing increasingly popular in the social sciences. Instead of being constrained by conventional wisdom and the myth of the 'perfect experiment', researchers can take advantage of existing *practical* solutions and develop creative alternatives for ethical and politically feasible implementations. In addition to the moral obligation to 'do no harm', neglecting these issues will almost certainly result in project failure and problems for future generations of experimentalists.

References

- BANERJEE, ABHIJIT V. AND ESTHER DUFLO (2009), 'The Experimental Approach to Development Economics', *Annual Review of Economics*, 1(1): 151-78.
- BEAUCHAMP, TOM L., RUTH R. FADEN, R. JAY WALLACE, JR. AND LEROY WALTERS. eds. (1982) *Ethical Issues in Social Science Research*. Baltimore: The Johns Hopkins University Press.
- BEECHER, HENRY K. (1971), 'Experiments on Children', *The Lancet*, 297(7710): 1181.
- BLUSTEIN, JAN (2005), 'Toward a More Public Discussion of the Ethics of Federal Social Program Evaluation', *Journal of Policy Analysis and Management*, 24(4): 824-846.
- BOHM, PETER (1984), 'Are There Practical Demand Revealing Mechanisms', [online] Available from: <http://karlan.yale.edu/fieldexperiments/pdf/AreTherePracticable.pdf> [Accessed 9 April 2012].
- BORUCH, ROBERT F. AND JOE S. CECIL. eds. (1983) *Solutions to Ethical and Legal Problems in Social Research*. New York: Academic Press.
- BORUCH, ROBERT F. AND WERNER WOTHKE (1985), 'Seven kinds of randomization plans for designing field experiments', *New Directions for Program Evaluation*, 1985(28): 95-113.
- BOUDREAU, CHERYL AND MATHEW D. MCCUBBINS (2010), 'The Blind Leading the Blind: Who Gets Polling Information and Does it Improve Decisions?', *The Journal of Politics*, 72(2): 513-527.
- BURTLESS, GARY AND LARRY ORR (1986), 'Are Classical Experiments needed for Manpower Policy?', *Journal of Human Resources*, 21(4): 606-639.
- BUTLER, DANIEL M. AND DAVID E. BROCKMAN (2011), 'Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators', *American Journal of Political Science*, 55(3): 463-477.
- CAMPBELL, DONALD (1969), 'Reforms as Experiments', *The ANNALS of the American Academy of Political and Social Science*, 628(1): 47-58.
- CAMPBELL, DONALD (1974), 'Randomization', in Riecken and Boruch eds. *Social experimentation: A method for planning and evaluating social programs*. New York: Academic Press pp.
- CAMPBELL, SIOBHAN (2003), The feasibility of conducting an RCT at HMP Grendon. Home Office Online Report. <http://rds.homeoffice.gov.uk/rds/pdfs2/rdsolr0303.pdf> (accessed March 18 2012).

- CORSI, JEROME R.. (1983) Randomization and Consent in the New Mexico Teleconferencing Experiment: Legal and Ethical Considerations. In: Boruch, Robert F. and Joe S. Cecil ed. *Solutions to Ethical and Legal Problems in Social Research*, New York, Academic Press, pp. 159-170.
- COTTERILL AND RICHARDSON (2010), 'Expanding the Use of Experiments on Civic Behavior: Experiments with Local Government as a Research Partner', *The ANNALS of the American Academy of Political and Social Science*, 628(1): 148-164
- DEE, THOMAS AND BENJAMIN KEYS (2004), 'Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment', *Journal of Policy Analysis and Management*, 23(3): 471-88.
- DOOLITTLE, FRED AND LINDA TRAEGER (1990), *Implementing the National NTPA Study*. New York: Manpower Demonstration Research Corporation, p. 34.
- DRUCKMAN, JAMES N., DONALD P. GREEN, JAMES H. KUKLINSKI AND ARTHUR LUPIA (2006), 'The Growth and Development of Experimental Research in Political Science', *American Political Science Review*, 100(4): 627-635.
- EDSALL, GEOFFREY (1971), 'Experiments at Willowbrook', *The Lancet*, 2(7715): 95.
- FAIRWEATHER, GEORGE WILLIAM AND LOUIS G. TORNATZKY (1977), *Experimental Methods for Social Policy Research*, Elmsford, NY, Pergamon Press.
- FARRINGTON, DAVID P. AND DARRICK JOLLIFFE (2002), A feasibility study into using a randomised controlled trial to evaluate treatment pilots at HMP Whitemoor. Home Office Online Report 14/02. London: Home Office. www.homeoffice.gov.uk/rds/pdfs2/rdsolr1402.pdf (accessed March 18, 2012)
- GERBER, ALAN S., DONALD P. GREEN AND CHRISTOPHER W. LARIMER (2008), 'Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment', *American Political Science Review*, 102(1): 33-48.
- GOLDBY, STEPHEN (1971), 'Experiments at the Willowbrook State School', *The Lancet*, 1(7702): 749.
- GREEN, DONALD P. AND ALAN S. GERBER (2003), 'The Underprovision of Experiments in Political Science', *The ANNALS of the American Academy of Political and Social Science*, 589(1): 94-112.
- GREENBERG, DAVID H., DONNA LINKSZ AND MARVIN MANDELL (2003), *Social Experimentation and Public Policymaking: Evaluating the Moving to Opportunity and Social Experiment*, Washington, DC, The Urban Institute Press.
- HECKMAN, JAMES J. AND JEFFREY A. SMITH (1995), 'Assessing the Case for Social Experiments', *The Journal of Economic Perspectives*, 9(2): 85-110.
- HILLIS, JAY W. AND CAMILLE B. WORTMAN (1976), 'Some Determinants of Public Acceptance of Randomized Control Group Experimental Designs', *Sociometry*, 39(2): 91-96.

- HUMPHREYS, MACARTAN (2011), 'Ethical Challenges of Embedded Experimentation', *Comparative Democratization*, 9(3): 23-29.
- KING, GARY, EMMANUELA GAKIDOU, NIRMALA RAVISHANKAR, RYAN T. MOORE, JASON LAKIN, MANETT VARGAS, MARTHA MARÍA TÉLLEZ-ROJO, JUAN EUGENIO HERNÁNDEZ ÁVILA, MAURICIO HERNÁNDEZ ÁVILE 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): 479-506.
- KRUEGER, ALAN B. (1999), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics*, 114(2): 497-532.
- KRUGMAN, SAUL (1971), 'Experiments at the Willowbrook State School', *The Lancet*, 1(7702): 966.
- KRUGMAN, SAUL, JOAN P. GILES AND JACK HAMMOND (1967), 'Infectious Hepatitis: Evidence for Two Distinctive Clinical, Epidemiological, and Immunological Types of Infection', *The Journal of the American Medical Association*, 200(5): 95-103.
- LAM, JULIE, STEPHANIE WILSON HARTWELL AND JAMES F. JEKEL (1994), "'I prayed real ard, so I know I'll get in": Living with randomization', *New Directions for Program Evaluation*, 1994(63): 55-66.
- LEIGH, ANDREW (2003), 'Randomised Policy Trials', *Agenda*, 10(4):341-354.
- LEVITT, STEPHEN AND JOHN A. LIST (2009), 'Field experiments in economics: The past, the present, and the future', *European Economic Review*, 53(1): 1-18.
- LIDZ, CHARLES W. AND LOREN H. ROTH. (1983) The Signed Form-Informed Consent?. In: Boruch, Robert F. and Joe S. Cecil ed. *Solutions to Ethical and Legal Problems in Social Research*, New York, Academic Press, pp. 145-158.
- LOEWEN, PETER JOHN, DANIEL RUBENSON AND LEONARD WANTCHEKON. (2010) 'Help me Help You: Conducting Field Experiments with Political Elites', *The ANNALS of the American Academy of Political and Social Science*, 628(1): 165-175.
- MARGETTS, HELEN AND GERRY STOKER. (2010) The Experimental Method: Prospects for Laboratory and Field Studies. In: Marsh, David and Gerry Stokes ed. *Theory and Methods in Political Science*, New York, Palgrave Macmillan, pp. 308-24.
- MIKÉ, VALERIE (1990), 'Suspended Judgement', *Controlled Clinical Trials*, 11(3): 153-56.
- MIKÉ, VALERIE (1989), 'Philosophers Assess Randomized Clinical Trials: The Need for Dialogue', *Controlled Clinical Trials*, 10(3): 244-53.
- MORTON, REBECCA AND KENNETH C. WILLIAMS. (2008) Experimentation in Political Science. In: Box-Steffensmeier, Janet M., Henry E. Brady and David Collier ed. *The Oxford Handbook of Political Methodology*, New York, Oxford University Press, pp. 339-356.

- MORTON, REBECCA AND KENNETH C. WILLIAMS. (2010) *Experimental Political Science and the Study of Causality*. New York: Cambridge University Press.
- NICKERSON, DAVID W. (2005), 'Scalable Protocols Offer Efficient Design for Field Experiments', *Political Analysis*, 13(3): 233-52.
- OAKES, J. MICHAEL (2002), 'Risks and Wrongs in Social Science Research: An Evaluator's Guide to the IRB', *Evaluation Review*, 26(5): 443-479.
- OLKEN, BENJAMIN (2010), 'Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia', *American Political Science Review*, 104(2): 243-267.
- ORR, LARRY L. . (1999) *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks: Sage Publications.
- PALMER, TED AND ANTHONY PETROSINO (2003), 'The "Experimenting Agency": The California Youth Authority Research Division', *Evaluation Review*, 27(3): 228-266.
- PALUCK, ELIZABETH AND DONALD P. GREEN (2009), 'Deference, Dissent and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda', *Evaluation Review*, 27(3): 228-266.
- PAPPWORTH, MAURICE HENRY (1968), *Human Guinea Pigs: Experimentation on Man*, London: Routledge.
- PAPPWORTH, MAURICE HENRY (1971), 'The Willowbrook Experiments', *The Lancet*, 297(7710): 1181.
- PETROSINO, ANTHONY, CAROLYN TURPIN-PETROSINO AND JOHN BUEHLER (2003), 'Scared Straight and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency: A Systematic Review of the Randomized Experimental Evidence', *Annals of the American Academy of Political and Social Science*, 589(1): 41-61.
- ROTH, ALVIN E. (1995), 'Introduction to Experimental Economics,' in *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press.
- RIECKEN, HENRY W. AND ROBERT F. BORUCH Eds. (1974), *Social experimentation: A method for planning and evaluating social programs*. New York: Academic Press
- SHADISH, WILLIAM R., THOMAS D. COOK AND DONALD T. CAMPBELL. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
- SHAPIRO, SAMUEL (1971), 'Experiments at the Willowbrook State School', *The Lancet*, 1(7702): 966.
- SONDHEIMER, RACHEL MILSTEIN AND DONALD P. GREEN (2010), 'Using Experiments to Estimate the Effects of Education on Voter Turnout', *American Journal of Political Science*, 54(1): 174-189.

TEELE, DAWN. (*forthcoming*) Experimental Reasoning in Social Science. In: Teele, Dawn L. ed. *Field Experiments and their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, New Haven, Yale University Press, pp. 12-26.

'SOCIAL SCIENCE: TRY IT AND SEE', *The Economist* [online] Available from: <http://www.economist.com/node/1010802> [Accessed 9 April 2012].

TORP, HEGE, ODDBJØRN RAAUM, ERIK HÆS AND HARALD GOLDSTEIN. (1993) The First Norwegian Experiment. In: Jensen, Karsten and Per Kongshoj Madsen ed. *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives*, Copenhagen, The Ministry of Labour, pp. 97-140.

VERMEERSCH, CHRISTEL AND MICHAEL KREMER (2004), 'School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation', *World Bank Policy Research Working Paper No. 3523*

Review of *Field Experiments: Design, Analysis and Interpretation*, by Alan S. Gerber and Donald P. Green

Rocío Titiunik
University of Michigan
titiunik@umich.edu

As experiments are growing in many subfields of political science, it has become essential for scholars to agree on a convincing and transparent framework for its design and analysis. This goal will be more easily achievable with *Field Experiments* at hand.

When a treatment is randomized, the analysis stage can be very simple; after all, a simple difference-in-means between treated and control groups provides an unbiased estimate of the average treatment effect. But in practice, this is rarely sufficient to analyze experiments successfully. What happens, for example, when the difference-in-means is insignificant and researchers sequentially add covariates to a multivariate regression until the effect becomes significant? And how should one analyze an experiment in which units were assigned to treatment within covariate blocks? What about an experiment in which units interfere with each other after the treatment is assigned, or studies where units do not comply with the treatment they are assigned? And what if one is interested in the effects of mediating variables? Gerber and Green explain carefully how some of these complications can be avoided partially or entirely in the design stage of the experiment, while others can be addressed successfully during the analysis. Their book provides a clear, consistent, accessible and compelling framework to treat all aspects of experiments, from design to analysis to interpretation, in a unified way.

The book is particularly about field experiments, that is, experiments conducted in natural settings. Gerber and Green choose this focus because they consider field experiments to be superior to laboratory experiments in terms of the authenticity of their context, the treatments assigned, the participants involved and the outcomes studied. But, despite their preference for field experiments, all

the chapters in the book are applicable to any kind of experiment, with the exception of Appendix B (which suggests specific field experiments for class projects) and perhaps Chapter 12 (which discusses design challenges mostly specific to field experiments).

The manuscript is designed to be accessible to readers with just basic knowledge of statistics and econometrics. It will be an invaluable resource for anyone who wants to design and analyze their own experiments (or simply analyze an existing experiment), and an essential reference for novice and expert readers who want a unified treatment of basic concepts as well as recent developments in experimentation methods. *Field Experiments* is also ideally suited for teaching, since it illustrates its abstract concepts with a variety of examples, and has both conceptual and practical exercises at the end of each chapter. Gerber and Green's long reputation on voter turnout experiments may lead some readers to expect that the examples and exercises will be drawn primarily from their own previous work. This is not the case. They discuss applications in a wide range of substantive areas, including economics, political science, education, health and criminology – with worked examples in topics such as mandated representation, teachers' pay-per-performance, political candidate debates, school vouchers, and voter attitudes. It is easy to envision designing an entire course about experiments based on the book, or using the book to cover the experimental section of a more general causal inference or research design course.

Although *Field Experiments* is not explicitly divided in sections, the chapters fall naturally into three different categories. Chapters 2 through 4 introduce core concepts. This is where the reader first encounters a discussion of potential outcomes, the fundamental problem of causal inference, random assignment, independence, statistical inference, and covariate adjustment. The notation and concepts introduced in these chapters are used throughout the book in a consistent manner, one of the main reasons why the book is simple and accessible. Chapters 5 through 10 discuss complications or refinements over the ideal experimental model introduced in the first chapters: noncompliance, attrition, interference, heterogeneous treatment effects, and mediation. In the last chapters, the authors reflect on more general and practical aspects of field experiments, such as how to integrate research findings under a Bayesian framework, how decisions at the design stage may influence the analysis, and how to write a research proposal. Two final appendices discuss issues surrounding human subjects and suggest simple field experiments for class projects.

At its most general level, *Field Experiments* is about answering causal questions, not asking them. The causal question is the starting point, and the authors are agnostic about how these questions arise – they could be sparked by theories of human behavior, interest in the effects of policies, or just plain curiosity. They fully embrace the Rubin Causal Model, which is to say that they impose a clear restriction on what a cause can be: a cause is only a cause if each unit in the study can be potentially exposed or not exposed to it. Moreover, without random assignment “one cannot be sure whether the treatment and control groups are comparable because no one knows precisely why some subjects and not others came to receive the treatment” (p.8), which is why the authors (along with many, but not all, methodologists) give randomized experiments the highest place in the causal inference hierarchy. Gerber and Green's view about non-experimental research designs is pessimistic in nature, but from this it does not follow that *Field Experiments* should be ignored by scholars who do not normally study

experiments, quite the opposite. Many of the issues discussed (heterogeneous treatment effects, mediating variables, interference between units) certainly arise in non-experimental designs, making the book a valuable reference for this type of studies as well. More importantly, as the authors carefully discuss throughout the manuscript, even randomized experiments are limited in the ways in which they can address or solve these issues; unwelcome as it will be to some, the valuable lesson for non-experimental researchers is that they have an even harder task at hand, and the book explains exactly how and why.

This is particularly clear in Chapter 10, which discusses mediation analysis and is, in my view, the chapter where all the principles advocated by Gerber and Green come together. Undoubtedly, identifying the mediating mechanisms that channel the effect of a given treatment to the final outcome of interest is one of the ultimate goals of science, social and otherwise. It is also one of the most difficult. As the authors carefully discuss, the standard regression analysis of mediation leads to biased results unless strong assumptions are met; and these assumptions are not guaranteed to hold by the randomization of treatment. What's more, mediating effects cannot be obtained in general even when both the mediator and the treatment are randomly assigned. Gerber and Green use the potential outcomes framework successfully and compellingly to explain the reader exactly why this is the case: there is a fundamental mismatch between what an experiment can manipulate and the phenomenon that a mediating analysis seeks to understand. This is the point where the high bar set by Gerber and Green has stark consequences. Under the framework of causation that they have adopted, the conclusion is that mediation analysis, as conventionally construed, is simply not possible. The authors, who are committed to being consistent and principled, embrace this negative conclusion, but the chapter is nonetheless constructive and does not leave the reader empty-handed. Gerber and Green propose to conduct implicit mediation analysis instead, in which researchers rely on theory to manipulate mediators indirectly by adding and subtracting different elements of the intervention. For many ready readers, this will be seen as a proposal to adhere to the potential outcomes framework in a principled way, inform it with theory, and make robust statistical inferences under minimal assumptions. But others may be less willing to accept Gerber and Green's conclusion that conventional mediation analysis is impossible. These readers will have to go back to Chapters 2 and 3 and reject the core assumptions; otherwise, they risk being pushed into a corner. This fundamentally principled approach, present from the first to the last chapter, is what makes the book so compelling.

Another of the main contributions of the manuscript is that it adopts a randomization-inference approach, deriving p-values and confidence intervals directly from the random assignment of treatment. This is a powerful framework, because it allows researchers to derive the distribution of a large class of test statistics from a small number of basic assumptions. In its purest form, this approach has three core assumptions: (i) the units who participate in the experiment are fixed, as opposed to a random sample from a larger population; (ii) the treatment is randomly assigned; and (iii) the randomization distribution of the treatment is known. Since the only source of randomness in this framework is the assignment of treatment, for some null hypotheses it is possible to construct test statistics whose distribution can be entirely derived from the randomization distribution of treatment. This leads to exact p-values and confidence intervals that are correct for any sample size, and thus

avoids the need to rely on large-sample approximations or parametric assumptions.

Gerber and Green follow this model to a large extent, but introduce some modifications along the way. First, although they consider potential outcomes as fixed attributes of each subject and all their analyses and examples conceive the treatment assignment as the only source of randomness, some of the core concepts introduced in Chapters 2 and 3 (such as the average treatment effect and statistical independence) and later discussions require at times conceiving potential outcomes as random variables. Although this may be unsettling for randomization-inference “purists”, there is not necessarily a contradiction because, at least for permutation-invariant test-statistics, there is a way to recast the original randomization-based framework to reconcile the core notion of fixed potential outcomes with Gerber and Green’s notation (see footnote 4 in the book and the discussion in Samii and Aronow (2012)). Still, for those readers used to the randomization-inference framework presented, for example, in Rosenbaum (2002), the manuscript will invite some reflection on important conceptual distinctions. On the other hand, this issue will likely go unnoticed to non-methodologists.

The other modification is entirely practical. In randomization-based inference, confidence intervals for a constant treatment effect are typically obtained by inverting hypothesis tests, as opposed to from closed-form formulas as is the case in standard inference. This is because the quantiles of the randomization distribution of the test statistic are not known in advance (as they depend on the particular way in which the treatment was randomized). Conceptually, this is appealing and straightforward, but it is practically difficult to implement, since it requires doing a large number of hypothesis tests. Since each of these tests requires permutations or simulations, this can become computationally demanding – and not entirely accessible to novice readers. Gerber and Green propose a simplified version of this, which seems to work well in practice while preserving the *distribution-over-all-possible-treatment-assignments* intuition essential to randomization-based inference. This involves simply using the estimated average treatment effect as the stipulated true constant treatment effect, and adjusting observed outcomes accordingly to obtain the average treatment effect for all possible permutations of the treatment. The disadvantage of this simplification is that it may introduce some inaccuracies – although these will tend to be small, particularly when the sample size is large. But for this modest cost, their method requires a single set of permutations or simulations, and makes the randomization-inference framework more feasible and accessible.

Overall, *Field Experiments* is a decisively successful book. It provides an accessible, principled and rigorous framework to the design and analysis of experiments that will become a standard reference in experimental research. It also shows that once the premises of the potential outcomes framework are adopted, the road to answering causal questions is principled and straightforward. But I dare to say that the most valuable aspect of *Field Experiments* is that despite choosing experiments over observational studies and fields over labs, it has a broad message of analytic rigor, transparency, replicability and responsibility in quantitative research. Or more succinctly, it is a great gift to our discipline.

References

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer 2nd edition.

Samii, Cyrus and Peter M. Aronow. 2012. On equivalencies between design-based and regression-based variance estimators for randomized experiments." *Statistics and Probability Letters* 82 (2): 365-370.

Section News and Announcements

• The Harvard-IDC Symposium in Political Psychology and Decision Making

From Julia Elad-Strenger

In the past year, the Program in Political Psychology and Decision Making (POP-DM) at the Lauder School of Government, Diplomacy and Strategy at IDC-Herzliya, Israel, has hosted several events providing an arena for scholars to discuss state-of-the-art experimental research and methods within the field of political psychology and behavioral decision-making. Two of the key events this year were the first annual Harvard-IDC symposium on Political Psychology and Decision Making and the Columbia-IDC dialogue on Decision Making, Framing and Negotiation, both featuring some of the world's leading experts in their field.

In the first annual Harvard-IDC symposium, held in November 2011 at the IDC, James Sidanius of Harvard University presented recent experiments testing the theory of gendered prejudice. One of the key findings presented by Sidanius is that males in disadvantaged groups are perceived more negatively and thus suffer from higher discrimination than females from the same group. Alex Mintz, editor-in-chief of the international journal, *Political Psychology*, and Director of POP-DM at the IDC, presented a series of computerized process tracing experiments utilizing Decision Board software to analyze the decision making patterns of Israeli, Palestinian and American leaders. Based on these analyses, Mintz and his colleagues formulated a two-stage poliheuristic decision making model that combines elements of the rational and cognitive schools and can be used to explain and predict foreign policy and national security decisions. Ryan Enos of Harvard, presented experimental work conducted in Chicago testing the effects of racial threat and group proximity on voting turnout.

The Harvard-IDC symposium also featured Ronit Kampf of Tel Aviv University and Esra Cuhadar of Bilkent University who presented a cross-cultural experimental study demonstrating the effectiveness of simulations of the Israeli-Palestinian conflict as tools for knowledge acquisition and attitude change towards the conflict among Israeli-Jews, Palestinian, American and Turkish students. Ilanit Siman Tov-Nachlieli, Nurit Shnabel and Arie Nadler of Tel Aviv University presented an experimental project demonstrating the interaction between exposure to threats to Israel's power or morality and the need for 'powerful Israel' and 'moral Israel' on pro- and anti-social behaviors among Israeli-Jews. Gilad Hirschberger of the IDC reported a series of experimental studies suggesting that mortality salience (MS) increases support for intergroup violence for reasons pertaining to justice rather than utility

motivations. Pazit Ben-Num Bloom of the Hebrew University presented two experiments testing the extent to which moral and non-moral conviction and primed harm associations and disgust hinder attitude change in a persuasive context. Yechiel Klar and Hadas Baram of Tel Aviv University reported experimental studies employing cognitive and behavioral methods to examine the effects of exposure to historical narratives regarding the Israeli-Arab conflict on ego-depletion, executive failure and narrative recall among Israeli students. Ifat Maoz of the Hebrew University presented experimental findings on the effect of opponent-negotiators' gender on evaluation of compromise solutions in the Israeli-Palestinian conflict.

POP-DM is scheduled to host the second annual Harvard-Lauder symposium in November 2012. This symposium will feature Dustin Tingely of the department of Government at Harvard University, who will conduct a workshop on his experimental work, and Gary Orren and Jennifer Shkabatur of the Kennedy school at Harvard University who will present their ongoing projects on persuasion and electoral campaigns respectively.

In a recent Columbia-Lauder School joint dialogue, held at POP-DM, Alex Mintz presented his ongoing experimental project on the Polythink syndrome, in which decision matrices of political leaders and decision makers are analyzed for the number of overlapping alternatives and dimension sets, to uncover whether the group suffers from Groupthink or what Mintz and DeRouen (2010) call Polythink. Mintz applied this method in an analysis of delegates to the Israeli-Palestinian 2000 Camp David Summit. Mark Whitlock of Columbia University discussed his experimental project on the role of web-based simulations in conflict education. Other participants in the workshop included Peter Coleman and Beth Fisher from Columbia University.

POP-DM, now housing the journal Political Psychology, has also been selected by the Governing Council of the International Society of Political Psychology (ISPP) to host its 2013 international annual conference, which will bring together hundreds of leading scholars from various fields to the IDC, and will include multiple sessions featuring experimental work.

- **WCE 2012 Report**

From Kevin Esterling

The fifth annual meeting of the West Coast Experiments (WCE.2012) Conference was hosted by UC Berkeley at the beautiful Claremont Resort on May 11, 2012. The WCE conference is organized as a one-day methods workshop, where presenters focus on innovative methods for experimental design and analysis. WCE.2012 was a great success, with 70 participants (half faculty and half graduate students) and seven presentations.

The morning panel focused on research designs that could identify the causal effects of things that cannot be assigned. Claire Adida (UCSD) presented her paper on "Isolating the Ethnic Effect on Political Behavior: A Survey Experiment Using Mixed Ethnicity Leaders in Benin," and Pradeep Chhibber (UC Berkeley) and Jasjeet Sekhon (UC Berkeley) presented "Mobilizing Identities: Experiments in India and Afghanistan on Religious Practice and Political Behavior." Both of these

papers used innovative survey experiment designs, navigating difficult logistical challenges in the field. Daniel Butler (Yale) presented "Estimating the Causal Effects of Progressive Ambition" which used a natural experiment based on exogenous variation in eligibility to run for higher office to identify the effect of ambition on legislative behavior in the West Virginia state legislature.

The first afternoon panel gave us two papers on development aid field experiments in west Africa. Jim Fearon (Stanford) discussed the results found in his paper on "Democratic Institutions and Collective Action Capacity: Results from a Field Experiment in Post-Conflict Liberia," and then Edward Miguel (UC Berkeley) presented "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." The second afternoon panel presented two papers on credibility and political communication. Adam Berinsky (MIT) presented "Rumors, Truth, and Reality: A Study of Political Misinformation" and Justin Grimmer (Stanford) spoke on "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation."

Most of the papers are available at the conference [website](#). Please also check this website later this summer for information on next year's meeting, WCE.2013.

- **Existing Newsletter Sections Need Authors**

I need future authors!! Book reviews (see Gailmard's of Norton/Williams, nominate your own book and reviewer too!), theme sections (e.g., the bio-politics and GOTV sections), co-author/inter-disciplinary experiences....Submit! Next issue is November 2013

- **Call for book proposals**

Call for Proposals: Routledge Studies on Experimental Political Science www.routledge.com/politics

Series Editors: Kevin T. Arceneaux, Temple University and Cindy D. Kam, Vanderbilt University

Advisory Board: Howard Lavine, University of Minnesota; Joshua Tucker, New York University; Rick Wilson, Rice University; and Elizabeth Zechmeister, Vanderbilt University

Political scientists are increasingly using experiments to study important political and social phenomena. The logic of experimentation makes it an appealing and powerful methodological tool that enables scholars to establish causality and probe into the mechanisms underlying observable regularities. Experiments, because of their transparency, also enable researchers to communicate their findings to a broad audience. Although highly technical knowledge is not necessary for understanding the gist of experiments, experiments must be designed, administered, and analyzed with care and attention to detail.

The Routledge Studies on Experimental Political Science was developed to publish books that educate readers about the appropriate design and interpretation of experiments and books that showcase innovative and important applications of experimental work. We are particularly interested in scholarly monographs, but proposals for edited volumes will also be considered.

The series will showcase experimental work in political science in at least two ways:

Pedagogy: Books that provide pedagogical guidance on the design, administration, and analysis of experiments, particularly tailored for a political science audience. Such books would be targeted at advanced undergraduates, graduate students, and faculty members.

Applications: Books that use experimental methods, particularly innovative experimental methods, to understand important causal relationships in political and social settings.

If you have a book proposal or idea in mind which might be suitable for the series, please do not hesitate to contact:

Series Editors Kevin Arceneaux: arceneau@temple.edu Cindy Kam: cindy.d.kam@vanderbilt.edu

Acquisitions Editor Natalja Mortensen, Political Science Research:
natalja.mortensen@taylorandfrancis.com

APSA Experimental Section Officers

- President: Shanto Iyengar (2012)
- President-elect: Alan Gerber (2012)
- At-large Council: Andre Blais (2012-13), Jennifer Jerit (2012-13), Jamie Druckman (2012-13), Don Green (2011-12)
- Treasurer: Jennifer Merolla (2012)
- Secretary: Rikhil Bhavnani (2012)
- Newsletter editor: Dustin Tingley (2010-12)
- Webmaster: Kevin Esterling (2010-12)
- APSA Program Chair: Ken Williams (2012)
- Nominating Committee (2012): Costas Panagopoulos (chair), John Bullock, Rick Matland, Rose McDermott
- Outreach Committee (2012): Yanna Krupnikov (chair), Adam Levine, Spencer Piston, Monica Schneider, Cheryl Boudreau
- Standards Committee: Alan Gerber (chair), Kevin Arceneaux, Tom Palfrey, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus
- Journal Advisory Committee: John Geer (chair), Kevin Arceneaux, Macartan Humphreys, Tom Palfrey, and Lynn Vavreck.
- Best Paper at prior APSA: Daniel Rubenson (chair), Dan Myers, Chris Weber
- Best Dissertation in prior calendar year: Jason Reifler (chair), Elizabeth Suhay, Reuben Kline
- Best Book in prior calendar year: Melissa Michelson (chair), Adam Berinsky, Thad Dunning