Field Experiments and Behavioral Theories: Science and Ethics

Elizabeth Carlson, *Pennsylvania State University*

ABSTRACT The design of field experiments makes them inappropriate for testing many common political theories. These theories propose that certain factors—for example, income or information—affect how individuals make choices. To test these theories, researchers typically investigate the correlation between the relevant factor and individuals' choices, holding other factors constant. Field experiments, in contrast, allow multiple factors to vary at once: they create real-world disruption and do not control how actors behave in response. Subjects' choices will be affected by the experimental treatment as well as by other changes that occur as the larger system reacts. It will be difficult to isolate the effect of any one factor, particularly without a good preexisting model of the system and how it is likely to respond. If a field experiment will not tell us what we need to know, the benefit of the study cannot outweigh harm, and it also will be unethical.

Field experiments are now common in the political science literature. Many of these experiments are randomized controlled trials or program evaluations, in which researchers randomize interventions by governments or aid organizations to determine the effects of these policies on desired outcomes. However, in part due to a push to make field experiments theoretically as well as empirically valuable, and in part due to the perceived advantage of field experiments in delivering estimates that are both causally identified and externally valid, field experiments also are being used to test or draw conclusions about behavioral theories.

I argue that field experiments are generally inappropriate for testing behavioral theories. Behavioral theories propose that individuals take in inputs, use them to determine which attitude or behavior will give them the most utility, and act accordingly. They generally are tested by estimating the correlation between the input of interest and the subjects' behavior, holding all else constant with experimental or statistical control. Field experiments, in contrast, capture how outcomes change after experimenters create disequilibrium in systems of strategic actors and neither control how these actors respond nor isolate treated subjects from the rest of the system. The actual treatment to which subjects are exposed will be a bundle of the experimenter's intervention and the system's endogenous response to intervention. The average treatment-effect calculation does not control for endogenous response and therefore does not estimate the effect of the manipulated input, per se, on subjects' behavior.

Elizabeth Carlson *is assistant professor of political science and African studies at Pennsylvania State University. She can be reached at ecc13@psu.edu.*

In principle, researchers can isolate the direct impact of their intervention by anticipating, preventing, or accounting for other actors' strategic response. In some countries, for which there is an extensive existing literature about the behavior of citizens and elites, this may be something researchers can do credibly. For many countries in the world, however, there is no such literature and—as evidenced by the frequency with which field experiments generate unexpected results—researchers are missing critical information about who the relevant actors are and how they are likely to respond to intervention. Researchers working in these contexts will be limited in their ability to model endogenous response to treatment and, therefore, to infer a behavioral parameter from a field-experimental treatment effect.

The mismatch between behavioral theories and field-experimental approaches also has ethical implications. No risk of harm to subjects is justifiable if a study will not yield the knowledge it is intended to yield; by disrupting real social and political systems—often at large scale and without subjects' consent—field experiments pose a particularly high risk.[1] Researchers who want to use a field experiment to test a behavioral theory have an obligation to demonstrate *ex ante* that they will be able to extract an interpretable behavioral parameter from their result. This means that they also must demonstrate sufficient knowledge of the system to anticipate and account for endogenous response to the intervention. When they cannot do so, proceeding with a field experiment—rather than with an approach that provides more leverage or less risk—will be unethical.

A STYLIZED EXAMPLE

To illustrate how the design of a field experiment impedes inference about behavioral parameters, consider a straightforward

doi:10.1017/S1049096519001112

experiment intended to test a simple theory. Based on the principal–agent model, the experimenters propose that corrupt leaders retain their position because of an information asymmetry. Voters prefer to sanction corruption but cannot get accurate information about its extent, leaving corruption only weakly correlated with voters' observed choices. Therefore, the experimenters audit local leaders and disseminate the results in leaflets in randomly selected constituencies. To measure the effects of treatment, the experimenters conduct surveys and gather election returns.

Now suppose the analysis shows that the intervention had no effect: there was no difference across treatment and control, on average, in either the reported support for the incumbent or official electoral returns. What can we infer about voters' preferences over corruption or the presence of an information asymmetry?

One explanation for the null result is that the theory is wrong: there was no information asymmetry, voters are indifferent to corruption, or both. However, another explanation is that the theory is correct—voters both care about corruption and need information about it—but the treatment was counteracted by a strategic response from other actors. Leaders concerned about losing vote share following an incriminating audit might have contested the information, released equally damning information about the challenger, increased private handouts, or engaged in intimidation against those who received information. Assuming voters have more than one variable in their utility functions, any of these scenarios might have prevented treated subjects from reducing their stated or actual support for the incumbent—even if they were highly attentive to the information in the leaflet. From the null result, we can state confidently that the total effect of information, *plus* any other changes spurred by the release of that information, was zero: it does not logically follow that the effect of the information on voters' attitudes and behaviors was zero.

Interpretation is no less problematic if the experimenters found, as expected, that treated voters reduced their support for leaders exposed as corrupt. If the regime is concerned about a scandal, it may respond by removing its support from leaders who have been publicly outed as corrupt. The opposition, sensing a potential vulnerability, may spend more resources campaigning in areas treated with bad news about the incumbent. Any of these scenarios could reduce incumbents' vote share, even if treated voters were entirely indifferent to the leaflets. Again, we can estimate confidently that the total effect of information plus the system's response to that information was significantly negative, but we cannot conclude that the information itself had a significant effect on voters' attitudes or behaviors.

Although the experiment yields a well-identified empirical conclusion about the effect of publicly releasing audits on vote share, it can provide little evidence about the direct effect of either information or corruption on voters' utility or choices. If other political actors responded to counteract (or amplify) the information intervention—which they had a strong incentive to do—then treated subjects were exposed to a bundled treatment, and the treatment effect cannot be taken as the effect of the directly manipulated input.[2]

FIELD EXPERIMENTS AND BEHAVIORAL THEORIES

The concerns raised in this article are not relevant to all field experiments. The inferences from pure policy evaluations are not threatened—and are likely improved—by exposing subjects to a treatment that bundles intervention and the system's endogenous response to intervention. My argument pertains specifically to the use of field experiments for testing behavioral theories.

No risk of harm to subjects is justifiable if a study will not yield the knowledge it is intended to yield; by disrupting real social and political systems—often at large scale and without subjects' consent—field experiments pose a particularly high risk.

Defining Behavioral Theories

Behavioral theories are those that propose some input or a set of inputs affects the utility individuals gain from different behaviors and the frequency with which they make particular choices. Field experiments have been designed to test or taken as evidence for various behavioral theories. Examples include theories that:

- Civic knowledge increases the likelihood that citizens will engage with formal channels of participation (Mvukiyehe and Samii 2017; Sexton 2017).
- Exposure to social sanctions reduces free riding and increases participation (Ferree et al. 2018; Gerber, Green, and Larimer 2008).
- Interaction with members of an outgroup reduces discrimination (Broockman and Kalla 2016; Paluck 2010).
- Priming the moral content of voting reduces voters' susceptibility to vote buying (Larreguy et al. 2017; Vicente 2014).

Critically, implicit in all of these theories is a claim that the input affects the outcome, *all else equal*. In testing these theories, we control for other predictors of these behaviors either experimentally or statistically.

Defining Field Experiments

For purposes of this article, a defining characteristic of a field experiment is that it creates a real-world disequilibrium. The causal claims generated by field experiments rest on comparisons of outcomes in existing (i.e., control) and counterfactual (i.e., treatment) states of the world. Such a comparison requires that treatment conditions be different enough from existing conditions that treated actors will choose measurably different behavior than they choose under the status quo. Because actors will not change strategies in equilibrium, treated subjects must be disequilibrated.

The other relevant feature of a field experiment is that once equilibrium has been disrupted, researchers do not control what happens between intervention and measurement. Subjects' attempts to reach a new equilibrium after intervention will not be prevented from disequilibrating other actors whose utilities are influenced by subjects' behavior; field experimentalists do not

https://doi.org/10.1017/S1049096519001112 Published online by Cambridge University Press

isolate treated subjects from others with whom they are socially, economically, or politically connected.[3] These indirectly treated actors also may respond, re-treating treated subjects or newly treating members of the control.

This jumble of dynamic, exogenous, and endogenous effects is a key feature of field experiments, not a flaw. In some cases, triggering a response from the broader system is an explicit goal of the intervention: consider the common design in which treatment is assigned to voters but the measured outcome is the behavior of politicians.

Theory–Method Mismatch

It should be apparent that there is a mismatch between the estimates of interest in behavioral theory and the estimates returned by a field experiment. Behavioral theories theorize the effects of particular inputs on behavior; the estimate of interest is the coefficient on the variable (β_x). Field-experimental treatment effects, in contrast, estimate the difference in average outcomes ($\Delta\overline{Y}$) between treated and control. When nothing changes between treatment and control except the variable of interest (x), then these two quantities will be the same. However, because field experiments provide incomplete experimental control, they do not and cannot ensure that there will not be other conditions changing along with x.

Existing Studies

I do not know the proportion of existing field experimental estimates that are affected by uncontrolled confounding: it is difficult to show that a confounder exists when it has not been identified. However, a small but increasing number of studies provide evidence that experimental interventions trigger strategic response from other actors. Larreguy et al. (2017), for example, showed that an anti-vote-buying campaign in Uganda increased vote buying by opposition candidates, which may explain why treatment reduced support for the incumbent. Cruz, Keefer, and Labonne (2016) showed that Filipino voters given congenial information about challengers' policy priorities still voted for the incumbent because the incumbent targeted them with increased handouts.

It should be apparent that there is a mismatch between the estimates of interest in behavioral theory and the estimates returned by a field experiment.

In the terminology of causal inference, a field experiment allows violations of the exclusion restriction. Alternatively, the problem can be described as omitted-variable bias: we are interested in the impact of x on Y. However, Y also is affected by other variables, which—because they were a response to a shift in x—are correlated with x. Because field experiments neither model the process that produces Y nor control anything other than x, we should expect a field experiment to be mis-specified for estimating β_x in the same way we expect a bivariate regression to be mis-specified.

As with any method in which there is potential for confounding, there also are ways to prevent or account for it. There may be cases in which it is reasonable to argue that the relevant political system is simply the subject, and other actors either will be unaffected by the intervention or unable to respond in a meaningful way. In these cases, the intervention will be the only difference between treatment and control.[4] Similarly, just as we can for observational analysis in a multicausal system, we could control for other actors' interference and isolate β_x statistically.

Both fixes undermine some of the advantages of field experimentation. The first limits the range of theories that we can test to those in which individuals think and act in ways that do not meaningfully affect others' incentives. The second means that the accuracy of our causal estimate depends on the accuracy of our model, with all of the assumptions and limitations that that implies.

More important, both fixes also require substantial preexisting knowledge. To identify and control for the system's response to treatment, we need to know the relevant actors and their utility functions. Even to make the case that there will be no systemic response, we need to know who other actors might be and enough about their incentives to be confident that the subjects' response to treatment will not affect them. In many countries where field experiments are being used to test behavioral theory, the discipline is only just beginning to model actors' incentives and behaviors. Indeed, in many of these countries, researchers use field experiments to evaluate policy interventions precisely *because* they do not have an adequate model to predict the intervention's outcome *a priori*.

Two interventions intended to increase political participation in African countries instead reduced participation because the treatment also increased the rate at which treated subjects encountered political or social intimidation (Ferree et al. 2018; Gottlieb 2016).

It seems plausible that strategic interference with treatment also could explain other findings in the literature, including that accusations of corruption only harm challengers (de Figueiredo, Hidalgo, and Kasahara 2012); that reports on wrongdoing have no effect on vote choice in Brazil despite strong effects in the lab (Boas, Hidalgo, and Melo 2018); that discouraging vote buying increases support for the incumbent in Nigeria (Vicente 2014); and that disseminating information to the public improves local governance even where electoral accountability is weak (Bjorkman and Svensson 2009).[5]

Risk of Harm

Field experiments are not the only approach prone to uncontrolled confounding. However, when the risk of confounding undermines a field experiment's advantage in causal identification, it is difficult to justify either the substantial time and money often required for a large-scale field experiment or the increased risk of long-term, real-world harm to subjects.

As before, I cannot estimate the proportion of field experiments that have harmed subjects when the harm was neither anticipated nor measured. However, enough studies have recorded serious harms such that we should consider the *ex-ante* risk of harm to be high. In several previous examples, as well as others (e.g., Blattman, Hartman, and Blair 2014;

https://doi.org/10.1017/S1049096519001112 Published online by Cambridge University Press

Mvukiyehe and Samii 2017), intervention increased the rate at which subjects were exposed to violence. Paluck (2010) showed that community discussions in the Democratic Republic of the Congo increased grievance and reduced donations to an outgroup. Multiple studies find that cash transfers reduce welfare for control subjects because transfers cause local inflation in the price of food and other necessities (Filmer et al. 2018; Lehmann 2013).

Researchers who do not have sufficient information to anticipate harm also do not have sufficient information to anticipate and control for the systemic response that may be confounding their intervention.

My point is not that the cited studies were unethical: many were evaluations of policies that would have happened regardless and identifying the effects of these interventions was necessary and valuable. However, they nevertheless provide evidence that even expert researchers and their local partners can fail to correctly predict the outcome of their interventions. Researchers who do not have sufficient information to anticipate harm also do not have sufficient information to anticipate and control for the systemic response that may be confounding their intervention. When preexisting information is limited, it will be difficult for researchers interested in testing a behavioral theory to credibly argue that the risk posed by a field experiment is lower or that the value provided by a field experiment is higher than other approaches and, therefore, to satisfactorily argue that a field experimental approach is ethical.

CONCLUSION

The overall conclusion of this article is that there is a fundamental mismatch between what we often mean by theory—that is, explanations of how particular inputs affect actors' decisions—and what field experiments are designed to return. The lack of control inherent in a field experiment means that actors' endogenous responses to intervention may be bundled with the experimenters' intervention; unbundling will be difficult or impossible without a strong preexisting model of the system. Random assignment is sufficient to generate a causally identified treatment effect, but not necessarily an estimate of the effect of the input of interest.

This argument is not particularly novel. Many other scholars have noted the shortcomings of field experimentation for theory testing (e.g., Deaton and Cartwright 2018). In their handbook, Gerber and Green (2012) provided several examples in which strategic interference with treatment prevents inference about the effect of the intervention itself. Gerber and Green did not emphasize, however, how common strategic interference is likely to be when experiments are designed to induce measurable changes in real-world political outcomes, or how often political scientists ask the types of questions for which isolating the effect of particular input is crucial. They also failed to note that their solution—that is, careful design that anticipates and accounts for strategic response—will not be effective when researchers do not already know the system's relevant actors or their incentives.

Although certainly not every field experiment will be affected by confounding, there is nothing about field-experimental design that will prevent it and nothing about the average treatment-effect calculation that can control for it. Just as for any method that does not offer complete experimental control, we should start from the assumption that estimates are biased unless and until researchers convincingly demonstrate otherwise. When researchers cannot do this, a field experiment obviously does not provide more leverage than other approaches. Researchers have an ethical obligation to choose an approach that provides more control, less risk of harm, or both. ■

NOTES

1. Any interaction with human subjects can cause disruption, but other types of studies tend to enroll fewer people and often are easier to conduct without alerting interested outside actors. More important, subjects who participate in a survey or lab experiment know they are participating: if the study will have repercussions, they can discontinue their participation or alert the researcher to the risks. For more about the ethics of field experiments, see Desposato (2015).
2. The system's strategic response to intervention also may lead to treatment of those in the control: for example, constituents in the control might protest until their leaders also released audits. However, because they arise in response to the experimenter's (non-) intervention, the nature of the additional treatments bundled to the intervention will be different across treated and control subjects. It is not necessary for actors to be aware of the experiment for their response to be correlated with treatment assignment. They must be able only to observe the intervention where it is assigned.
3. Some studies that are described as field experiments do not meet this criterion because they measure outcomes immediately after subjects are exposed to stimulus and before the system can respond. This type of design does not sacrifice experimental control and does not pose the problems with interpretation described in this article. For my purposes, these studies are lab experiments, not field experiments.
4. Actors' responses may be more constrained in countries with strong rule of law. Actors also will be less likely to respond when they do not know about the intervention, which might happen if treatment involves private communication. However, even studies focused on communication in countries such as the United States have generated unanticipated systemic response.
5. In this latter case, an alternative explanation is that information was spread to more powerful actors (e.g., the central government), who sanctioned underperforming leaders. For example, Avis, Ferraz, and Finan (2016) found that reduced corruption in Brazil following public release of audits was driven more by politicians' fear of prosecution than by their fear of electoral sanctions.

REFERENCES

Avis, Eric, Claudio Ferraz, and Frederico Finan. 2016. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." National Bureau of Economic Research, Working Paper No. 22443.

Bjorkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics* 124 (2): 735–69.

Blattman, Christopher, Alexandra C. Hartman, and Robert A. Blair. 2014. "How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education." *American Political Science Review* 108 (1): 100–120.

Boas, Taylor, F. Daniel Hidalgo, and Marcus Melo. 2018. "Norms vs. Action: Why Voters Fail to Sanction Malfeasance in Brazil." Unpublished manuscript.

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352 (6282): 220–24.

Cruz, Cesi, Philip Keefer, and Julien Labonne. 2016. "Incumbent Advantage, Voter Information, and Vote Buying." Inter-American Development Bank, Working Paper Series No. IDB-WP-711.

de Figueiredo, Miguel F. P., F. Daniel Hidalgo, and Yuri Kasahara. 2012. "When Do Voters Punish Corrupt Politicians? Experimental Evidence from Brazil." Unpublished manuscript.

https://doi.org/10.1017/S1049096519001112 Published online by Cambridge University Press

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

Desposato, Scott. 2015. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals.* Routledge Studies in Experimental Political Science. New York: Taylor & Francis.

Ferree, Karen E., Danielle F. Jung, Robert A. Dowd, and Clark C. Gibson. 2018. "Election Ink and Turnout in a Partial Democracy." Available at https://doi.org/10.1017/S0007123418000121.

Filmer, Deon P., Jed Friedman, Eeshani Kandpal, and Junko Onishi. 2018. "General Equilibrium Effects of Targeted Cash Transfers: Nutrition Impacts on Non-Beneficiary Children." World Bank Policy, Working Paper No. WPS8377.

Gerber, Alan, and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York: W. W. Norton.

Gerber, Alan, Donald P. Green, and Christopher Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.

Gottlieb, Jessica. 2016. "Why Might Information Exacerbate the Gender Gap in Civic Participation? Evidence from Mali." *World Development* 86: 95–110.

Larreguy, Horacio, Benjamin Marx, Otis Reid, and Christopher Blattman. 2017. "A Market Equilibrium Approach to Reduce the Incidence of Vote Buying: Evidence from Uganda." Unpublished manuscript.

Lehmann, M. Christian. 2013. "(Adverse) General Equilibrium Effects of Cash Transfers." Unpublished manuscript.

Mvukiyehe, Eric, and Cyrus Samii. 2017. "Promoting Democracy in Fragile States: Field-Experimental Evidence from Liberia." *World Development* 95: 254–67.

Paluck, Elizabeth Levy. 2010. "Is It Better Not to Talk? Group Polarization, Extended Contact, and Perspective Taking in Eastern Democratic Republic of Congo." *Personality and Social Psychology Bulletin* 36 (9): 1170–85.

Sexton, Renard. 2017. "The Unintended Consequences of Bottom-Up Accountability: Evidence from a Field Experiment in Peru." Unpublished manuscript.

Vicente, Pedro C. 2014. "Is Vote Buying Effective? Evidence from a Field Experiment in West Africa." *The Economic Journal* 124 (574): F356–87.

https://doi.org/10.1017/S1049096519001112 Published online by Cambridge University Press