CHAPTER 4

Social Policy: Mechanism Experiments and Policy Evaluations[a]

W.J. Congdon[*,1], **J.R. Kling**[§,¶], **J. Ludwig**[¶,||], **S. Mullainathan**[¶,**]

*ideas42, New York, NY, United States
§Congressional Budget Office, Washington, DC, United States
¶NBER (National Bureau of Economic Research), Cambridge, MA, United States
||University of Chicago, Chicago, IL, United States
**Harvard University, Cambridge, MA, United States
[1]Corresponding author: E-mail: bill@ideas42.org

Contents

[a] This chapter was prepared for the *Handbook on Field Experiments*, edited by Abhijit Banerjee and Esther Duflo, and draws heavily from: Jens Ludwig, Jeffrey R. Kling and Sendhil Mullainathan, 2011, "Mechanism experiments and policy evaluations," *Journal of Economic Perspectives*, 25 (3): 17–38. For excellent research assistance we thank Laura Brinkman and Michael Reddy. We thank Nava Ashraf, David Autor, Iwan Barankay, Jon Baron, Howard Bloom, Lorenzo Casaburi, Philip Cook, Stefano DellaVigna, John DiNardo, Esther Duflo, Judy Gueron, Elbert Huang, Chad Jones, Lawrence Katz, Supreet Kaur, John List, Stephan Meier, David Moore, Steve Pischke, Harold Pollack, Dina Pomeranz, David Reiley, Frank Schilbach, Robert Solow, Timothy Taylor, and conference participants at the University of Pennsylvania's Wharton School of Business, the American Economic Association, and the NBER Handbook of Field Experiments Conference for helpful comments. For financial support, we thank the Russell Sage Foundation (through a visiting scholar award to Ludwig). Any errors and all opinions are our own. The views expressed here are those of the authors and should not be interpreted as those of the Congressional Budget Office.

Handbook of Economic Field Experiments, Volume 2
ISSN 2214-658X, http://dx.doi.org/10.1016/bs.hefe.2016.10.004

Abstract

Policymakers and researchers are increasingly interested in using experimental methods to inform the design of social policy. The most common approach, at least in developed countries, is to carry out large-scale randomized trials of the policies of interest, or what we call here *policy evaluations*. In this chapter, we argue that in some circumstances the best way to generate information about the policy of interest may be to test an intervention that is different from the policy being considered, but which can shed light on one or more key mechanisms through which that policy may operate. What we call *mechanism experiments* can help address the key external validity challenge that confronts all policy-oriented work in two ways. First, mechanism experiments sometimes generate more policy-relevant information per dollar of research funding than can policy evaluations, which in turn makes it more feasible to test how interventions work in different contexts. Second, mechanism experiments can also help improve our ability to forecast effects by learning more about the way in which local context moderates policy effects, or expand the set of policies for which we can forecast effects. We discuss how mechanism experiments and policy evaluations can complement one another, and provide examples from a range of social policy areas including health insurance, education, labor market policy, savings and retirement, housing, criminal justice, redistribution, and tax policy. Examples focus on the US context.

Keywords

Field experiment; Program evaluation; Randomized controlled trial

JEL Code

C90

1. INTRODUCTION

Randomized experiments have a long tradition of being used in the United States to test social policy interventions in the field, dating back to the social experimentation that began in the 1960s.[1] The use of field experiments to test social policies has accelerated in recent years. For example the US Department of Education in 2002 founded the Institute for Education Sciences with a primary focus on running experiments, with an annual budget that was $574 million in 2015 (US Department of Education, 2015). This trend has been spurred in part by numerous independent groups that promote policy experimentation.[2]

This trend toward ever-greater use of randomized field experiments has led to a vigorous debate within economics about the value of experimental methods for informing policy (e.g., Angrist and Pischke, 2009, 2010; Banerjee and Duflo, 2009;

[1] Gueron and Rolston (2013), along with the chapter in this volume by Gueron, provide an account of this early period in the development of randomized demonstration projects for social policy.

[2] Examples include the Campbell Collaboration, the Jameel Poverty Action Lab at MIT, the University of Chicago Urban Labs, the Lab for Economic Opportunity at Notre Dame University, and the Laura and John Arnold Foundation.

Deaton, 2010; Heckman, 2010; Imbens, 2010). There is little disagreement that a well-executed experimental test of a given policy carried out in a given context provides a strong claim to internal validity—differences in outcomes reflect the effects of the policy within the experimental sample itself. The debate instead focuses on concerns about external validity—that is, to what other settings can the result of a given field experiment be generalized.

In the area of social policy and in many other areas, this debate has often been framed as a choice between experimental and nonexperimental methods. But this ignores an important choice of how to employ experimental methods that we argue here deserves greater attention. Specifically, in this chapter we argue (and demonstrate through numerous examples) that—perhaps counter-intuitively—the best way to test a policy is not always to directly test the policy of interest. Greater use could be made of randomized field experiments that test mechanisms of action through which social policies are hypothesized to affect outcomes—what we call *mechanism experiments*—even if the interventions tested do not directly correspond to those policies we are interested in understanding.

An example may help to illustrate our argument. Suppose the US Department of Justice (DOJ) wanted to help local police chiefs decide whether to implement "broken windows" policing, which is based on the theory that police should pay more attention to enforcing minor crimes like graffiti or vandalism because they can serve as a "signal that no one cares," and thereby accelerate more serious forms of criminal behavior (Kelling and Wilson, 1982, p. 31). Suppose that there is no obviously exogenous source of variation in the implementation or intensity of broken windows policing across areas, which rules out the opportunity for a study of an existing "natural experiment" (Meyer, 1995; Angrist and Pischke, 2009). To an experimentally minded research economist, the most obvious next step would be for DOJ to choose a representative sample of cities, randomly assign half to receive broken windows policing, and carry out what we would call a traditional *policy evaluation*.

Now consider an alternative experiment: Buy a small fleet of used cars. Break the windows of half of them. Randomly select neighborhoods and park the cars there, and measure whether more serious crimes increase in response. While this might initially seem like a fanciful idea, this is basically the design that was used in a 1960s' study by the Stanford psychologist Philip Zimbardo (as described by Kelling and Wilson, 1982, p. 31). The same idea was used more recently by Keizer et al. (2008) who examined the effects of various forms of disorder (such as graffiti or illegal firecrackers exploding) and found substantially more litter and theft occurred when they created disorder. One can of course perform variants with other small crimes, or randomly select neighborhoods for the reduction of disorder such as clean-up of smashed liquor bottles, trash, and graffiti. This *mechanism experiment* does not test a policy: it directly tests the causal mechanism that underlies the broken windows policy.

Which type of experiment would be more useful for public policy? The underlying issue is partly one of staging. Suppose the mechanism experiment failed to find the causal mechanism operative. Would we even need to run a policy evaluation? If (and this is the key assumption) we are confident that a policy implementing broken windows policing would affect crime only by reducing disorder and were convinced that we had strong evidence that reducing disorder does not affect crime, then we could stop. Running the far cheaper mechanism experiment first serves as a valuable screen. Conversely, if the mechanism experiment found strong effects, we might run a policy evaluation to figure out how much disorder could be reduced by applying broken windows policing at different levels of intensity. Indeed, depending on the costs of the policy evaluation, the magnitudes found in the mechanism experiment, and what else we think we already know about the policing and crime "production functions," we may even choose to adopt the policy straightaway.

In our example, mechanism experiments help us stretch research funding further, which bears directly on the ability of experimentation to create generalizable knowledge that is useful for social policy. One way to address external validity with randomized field experiments is replication—that is, testing the policy in many different contexts. As Angrist and Pischke (2010, pp. 23, 24) argue, "a constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge … the process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general."[3] One challenge to this strategy stems from resource constraints. In the broken windows application, mechanism experiments help with these resource constraints by incorporating prior knowledge and letting us focus on the issues about which the most remains to be learned.

In the spirit of contributing to a handbook that is intended to be of practical use to both policymakers and researchers, we organize the remainder of this chapter into three sets of applied questions: In Section 2 we clarify and expand on the answer to the question: What are mechanism experiments? Mechanism experiments can be defined broadly as tests of the causal mechanisms (M) that link policies (P) to social outcomes (Y). Mechanism experiments test the M → Y relationship using interventions that do not necessarily correspond to the actual policies of immediate interest. The connection of a mechanism to a clearly specified policy, not just to social outcomes, helps distinguish what we mean by mechanism experiments from more general efforts within economics to understand what determines outcomes.

In Section 3 we answer the question: Why do mechanism experiments? A primary motivation, as noted above, is to help establish external validity. Mechanism experiments can do this in two ways. First, they can increase the amount of policy-relevant information

[3] As Cook and Campbell (1979) note, "tests of the extent to which one can generalize across various kinds of persons, settings and times are, in essence, tests of statistical interactions … In the last analysis, external validity … is a matter of replication" (p. 73, 78).

per research dollar available, since replication of policy evaluations is a costly way to learn about external validity. Mechanism experiments can concentrate resources on parameters where policymakers have the most uncertainty, as in the broken windows example, or help us rule out policy evaluations that we don't need to run, or for that matter rule out policies, with the added benefit of sometimes reducing the amount of time required to realize that some policy is not actually promising. Second, mechanism experiments address questions of external validity related to forecasting the contexts in which a policy would have effects. Mechanism experiments can improve our ability to forecast effects by learning more about the way in which local context moderates policy effects, or expand the set of policies for which we can forecast effects.

In Section 4, we answer the questions: When should we do a mechanism experiment, when should we do a policy evaluation, and when should we do both? One necessary condition for doing a mechanism experiment is that researchers or policymakers need to believe they know at least something about the candidate mechanisms through which a policy affects social welfare. If the list of candidate mechanisms is short and the costs of carrying out a full-blown policy evaluation are high (or if the policy stakes are low), a mechanism experiment by itself might be sufficient to inform policy. Likely to be more common are situations in which it makes sense to follow a mechanism experiment with a policy evaluation to understand other links in the causal chain from policy to outcomes, or to calibrate magnitudes. The mechanism experiment still adds great value in these cases by helping us prioritize resources for those areas where a full-blown policy evaluation is worth doing. We note that in some situations, such as when there is a long list of candidate mechanisms that could have interactive effects or even work at cross-purposes, it may not be worth doing a mechanism experiment and researchers should just proceed to carry out a black-box policy evaluation.

While our discussion largely focuses on the key conceptual points behind our argument, we also try to illustrate the potential contributions (and limitations) of mechanism experiments with existing social policy studies whenever possible. As we discuss further below, at present mechanism, experiments are relatively more common in developing than developed country contexts, partly because for a variety of reasons development experiments are more typically carried out with NGOs than with government partners. The potential gains from rebalancing the policy experiment portfolio to include more mechanism experiments, not just policy evaluations, seem largest within the developed country context. Partly for that reason, we focus most of our discussion and examples on the developed country context with which we are most familiar ourselves, the United States. For a more comprehensive summary of social policy experiments that have been carried out in the United States, see Greenberg and Shroder (2004).[4]

[4] An updated version of their publication The Digest of Social Experiments is in progress.

2. WHAT ARE MECHANISM EXPERIMENTS?

2.1 Definition

Broadly, a mechanism experiment is an experiment that tests a mechanism—that is, it tests not the effects of variation in policy parameters themselves, directly, but the effects of variation in an intermediate link in the causal chain that connects (or is hypothesized to connect) a policy to an outcome. That is, where there is a specified policy that has candidate mechanisms that affect an outcome of policy concern, the mechanism experiment tests one or more of those mechanisms. There can be one or more mechanisms that link the policy to the outcome, which could operate in parallel (for example when there are multiple potential mediating channels through which a policy could change outcomes) or sequentially (if for example some mechanisms affect take-up or implementation fidelity). The central idea is that the mechanism experiment is intended to be informative about some policy but does not involve a test of that policy directly.

In our broken windows example, given above, one could sketch this model as follows: Policing policies (P) that target and reduce minor offenses such as broken windows (M) ultimately lead to reductions in the thing policymakers are most concerned about serious criminal offenses (Y) (Fig. 1). The hypothetical policy evaluation in that case—randomly assign cities to receive broken windows policing, and track outcomes for more serious crimes—is a policy evaluation, a test of P → Y. The result tells policymakers whether that policy has an impact on the outcomes of key policy interest. The corresponding mechanism experiment—randomly assign cars with broken windows across neighborhoods—is a test of M → Y. It tells policymakers whether the mechanism is operative.

Even though the mechanism experiment does not resemble in any way the policy of interest, it can concentrate resources on estimating the parameters most relevant to policy decisions, leading the experiment to be informative for policy to a surprising degree. Suppose that from previous work policymakers also know the elasticity of minor offenses with respect to policing (P → M), and they also believe that change in minor offenses is the only mechanism through which broken windows may affect the outcome of ultimate policy concern (serious offenses). What policymakers do not know is the accelerator: by how much will reducing minor offenses cascade into reducing serious offenses. The mechanism experiment estimates the parameter about

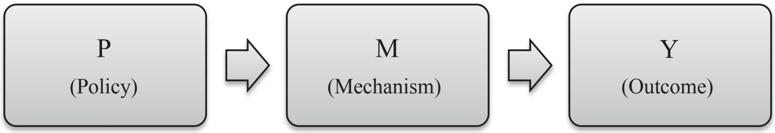

Figure 1 Policies, mechanisms, and outcomes.

which there is the greatest uncertainty or disagreement (M → Y). In contrast, the policy evaluation that measures the policy's impact on serious crimes, P → Y, also provides information about the crime accelerator, but with more noise because it combines the variability in crime outcomes with the variability in the impact of policing on minor crimes in any given combination of cities and years. With enough sample (that is, money) one could recover the M → Y link.

The broken windows example is not an isolated case: many policies have theories built into them, even if they are sometimes left implicit. Often these theories can be tested more cost effectively and precisely with experiments that do not mimic the policy of direct interest, or in some cases do not even mimic anything likely to ever be a real (or even feasible) policy. But while social scientists already value mechanism experiments because they contribute to building knowledge, we argue that *even if the sole goal were informing policy* (social policies or others), mechanism experiments, even those that do not test real policy options, play a crucial and under-appreciated role.

In our broken windows example, it is of course possible that a policymaker who is initially interested in the question of whether broken windows policing works sees the results of a mechanism experiment and decides that a policy to (say) remediate blight in distressed neighborhoods might itself be worth supporting at scale. The mechanism experiment in this case has the effect of reorienting policymaker attention to expand the set of policies (P's) they consider. But the key definitional point is: What we would consider a mechanism experiment depends on the initial policy question. Our goal is to help policymakers, policy-research funders and policy-oriented researchers see that if the focus is on broken windows policing, the experiment that yields the most useful information per dollar spent relevant to broken windows policing does not necessarily involve the delivery of broken windows policing.

Note that this definition does not imply that mechanism experiments are necessarily "cheap" in an absolute sense, only in a relative sense compared to other policy evaluations that would produce the same amount of policy-relevant information. For example, the RAND health insurance experiment was a mechanism experiment by our definition, since it was intended to tell us something about how cost-sharing provisions of health insurance plans affect health but included a treatment arm that did not correspond to a real (or even feasible) policy—a condition with zero co-insurance (Newhouse and the Insurance Experiment Group, 1993). Yet as we will argue below, this "extreme dosage" arm makes it possible for policymakers to interpolate the effects of plans with a wide range of co-insurance rates at cost lower than would be required to carry out at-scale policy evaluations of plans at each possible co-insurance rate. At a total cost of $285 million in 2010 dollars, the RAND experiment also holds the record—for now—as the most expensive mechanism experiment ever (Greenberg and Shroder, 2004, p. 181).

2.2 Some history

The use of mechanism experiments is not new in the US context, even if the label has been developed only relatively recently (Ludwig et al., 2011b). In fact, some of the earliest social policy experiments in the United States took the form of what we would classify as mechanism. For example, the New Jersey Income Maintenance Experiment is often considered the first large-scale social policy experiment in the United States (Greenberg and Shroder, 2004). This experiment was explicitly concerned more with testing mechanisms (behavioral responses to different marginal tax rates, including negative tax rates) than with the study of existing policies. This experiment and the RAND health insurance experiment mentioned previously were not isolated examples (even if they were not labeled as mechanism experiments at the time they were carried out).

In fact, a critique some economists leveled against the large-scale government social experiments of the 1970 and 1980s was precisely that they were "not necessarily test[ing] real policy options" (Harris, 1985, p. 161). In addition, budget changes in the 1980s reduced support for policy-relevant social experimentation of any sort. For these and other reasons, mechanism experiments of this sort fell out of favor in the United States over the ensuing decades.

Mechanism experiments continue to be relatively common at present in the developing world context, which has seen a substantial growth in policy-oriented experimentation over the past decade (Banerjee and Duflo, 2009). Of course there are large-scale policy evaluations that occur in developing country contexts, such as the *Progresa* conditional cash transfer experiment in Mexico (see for example Skoufias et al., 2001; Rivera et al., 2004; Schultz, 2004). There are also investigator-initiated policy evaluations, such as tests of ordeal mechanisms for receipt of social program assistance (Alatas et al., 2013). But the ratio of mechanism experiments to policy evaluations is higher in developing than developed countries.

One candidate explanation is that public sector capacity is typically more limited in developing countries, and so development experiments tend to be carried out frequently in partnership with NGOs rather than with government agencies. For example, Ashraf et al. (2010) study the role of prices in mediating the use of social services by sending privately hired marketers out to households to offer them a water chlorination product at different offered prices that are then transacted at a second randomized price. The policy of interest is the price at which the product is sold to households, which can affect usage and hence health (Y) through two candidate mechanisms—a screening mechanism (M_1) that affects which households use the product at different prices, and a sunk-cost mechanism (M_2) that suggests families may be more willing to use the product if they have paid more for it. The double randomization of the initial offer price and ultimate transaction price leads to an intervention that does not exactly correspond to an actual policy, but which does help separate out these two mechanisms and so help shed light on the optimal pricing and subsidy policies (in the same spirit also see for example Karlan and Zinman, 2009; Cohen and Dupas, 2010).

Our focus on the rest of the chapter will be on the developed world context, particularly the developed country we know best (the United States), since such a large share of the randomized experiments carried out in the United States that are intended to be helpful to social policy involve directly testing actual policies, often carried out in collaboration with government agencies. To take just one recent example, the experimental evaluation of the Reemployment and Eligibility Assessment (REA) initiative funded by the US Department of Labor was a large-scale test of a policy already in place in most states (Poe-Yamagata et al., 2011). The US Department of Education's Institute for Education Sciences (IES) describes its goals in funding experiments to be "development of practical solutions for education from the earliest design stages through pilot studies and rigorous testing at scale," and "large-scale evaluations of federal education programs and policies." Numerous private research firms around the country are carrying out multiple multimillion dollar policy evaluations at any point in time with support from private foundations and government. The resources and attention devoted today in the United States and other developed nations to what we could consider mechanism experiments pales in comparison to what is devoted to policy evaluations. In the next section, we develop our argument for why we believe this balance should shift over time.

3. WHY DO MECHANISM EXPERIMENTS?

In a given setting, which policy P generates the greatest change in outcomes Y at a given cost? That is the central question of primary interest to policymakers. Given that objective, why carry out mechanism experiments, which do not test actual (or perhaps even feasible) policies?

The answers are motivated partly by the inevitable question we have with any policy evaluation, which has to do with its external validity. The effects of, say, broken windows policing in Chicago's affluent North Shore suburb of Evanston may differ from what would happen as a result of this intervention in the distressed neighborhoods of Chicago's south side. We are worried that the "treatments" we study may interact with attributes of the target population, context, or time period. These baseline attributes that interact with treatment effects are what non-economists call "moderators."

This concern that treatments may interact with context has led naturally to the view that the best way to produce generalizable information is to focus on policy interventions of the sort that policymakers might actually implement, and test them in multiple settings of the sort in which the policy might actually be implemented. One way to think about what we are trying to accomplish through this replication comes from the useful distinction suggested by Wolpin (2007) and Todd and Wolpin (2008) between *ex-post policy* *evaluation*—understanding what happened as the result of a policy or program that was actually implemented—and *ex-ante policy evaluation*, which, as DiNardo and Lee (2011, p. 2) describe it, "begins with an explicit understanding that the program that was actually run may not be the one that corresponds to a particular policy of interest. Here, the goal is not descriptive, but instead predictive. What would be the impact if we expanded eligibility of the program? What would the effects of a similar program be if it were run at a national (as opposed to a local) level? Or what if the program were run today (as opposed to 20 years ago)? It is essentially a problem of forecasting or extrapolating, with the goal of achieving a high degree of external validity."

Replicating tests of real policies in different contexts tells us something about the value of using the policy's average effect as a forecast for what we would expect to accomplish in other settings. An obvious challenge with this approach is that policy evaluations are expensive and often difficult to carry out. One use of mechanism experiments is to increase the policy-relevant information we obtain for a given research budget, to maximize the coverage of policy-relevant contexts about which we have some information. Mechanism experiments help us do this by:

- Concentrating resources on parameters where there is the most uncertainty,
- Ruling out policies (and the need for full-blown policy evaluations), and
- Extracting more information from other (nonexperimental) sources of evidence.

Of course, evidence of treatment effect heterogeneity is not fatal to the idea of using policy experiments to help inform policy, since it is always possible to use forecasting methods that emphasize results from settings that are similar to whatever local context is being considered for some new policy. For example, we might predict the effects of broken windows policing in south side Chicago by focusing on results from Evanston's poorer neighborhoods specifically. Forecasting becomes essentially a matching or reweighting exercise (see, for example, Hotz et al., 2005; Cole and Stuart, 2010; Imbens, 2010; Stuart et al., 2011). The value of replicating tests of real policies comes from the fact that the chances of finding a "match" for any future policy application increases with the number of policy-relevant contexts in which the actual policy has been tested. Mechanism experiments can generate useful information for this type of policy forecasting exercise in two main ways:

- Understanding mechanisms of action can help shed light on those contextual factors that moderate policy effects, and so help forecast policy effects to different contexts, and
- Expanding the set of policies for which we can forecast policy effects by testing extreme policy "doses," so that forecasting relies on interpolation not extrapolation.

In the remainder of this section, we discuss these different uses for mechanism experiments in greater detail and include several examples. Because mechanism experiments remain underutilized in developed-country contexts, we present several hypothetical examples that illustrate the potential value-added of this approach for the study of social problems in places like the United States. Where possible we also present real examples of what we consider to be mechanism experiments, even if the authors themselves might not have explicitly set out to execute a mechanism experiment when they launched their studies.

3.1 Concentrating resources on the parameters with the most uncertainty

Social policymaking invariably involves operating along a causal chain of variable length. Policy reforms are reflected in statutory or regulatory changes, leading to corresponding adjustments in program administration and implementation, to which individuals or other actors respond along sometimes multiple margins. The result of what happens along the full causal chain is what ultimately ends up determining social welfare impacts. At each step, the impacts are uncertain, especially to the extent that ultimate impacts depend on behavioral responses.

Of course the option of testing the entire chain jointly through a full policy evaluation is always available. But depending on what we already know about some of the links in the chain, this might not be the most *efficient* way to learn about the likely effects of a policy. Suppose there is a policy (P) that is thought to affect some outcome (Y) through a candidate mediator or mechanism (M), so that the theory of change behind the intervention is expressed as: $P \rightarrow M \rightarrow Y$. If we already believe we understand the $P \rightarrow M$ link, for instance, we can concentrate resources on understanding the remaining part of the theory of change for the policy without having to incur the costs of a full policy experiment.

In this way, we can use field experiments to learn about social policy design without necessarily testing actual policies. A mechanism experiment will allow us to identify the response of Y to M, and, in combination with what we already believed we knew about $P \rightarrow M$, this allows us to learn what we ultimately want to know—about $P \rightarrow Y$—without having to test the full policy or every point in the logic chain. Under the most straightforward conditions—there is only a single candidate M, and the relationship between $M \rightarrow Y$ is stable—this boils down to, essentially: we can learn about the sign of the response of Y to M, the magnitude of the response of Y to M, and the shape of the response of Y to M.

For example, there is a great deal of concern right now in the United States about the quality of police-community relations, and in particular with the prevalence of use of force by police officers against civilians (particularly in high-crime, economically and racially segregated communities). One hypothesis for this use of force is officer exhaustion—many police officers wind up, working second jobs to make ends meet—which, combined with long and sometimes unpredictable hours at their main job, can lead to overtired officers whose ability to cope with stressful situations is impaired. A potential policy response would be to substantially increase police salaries. Suppose we were confident that big pay raises for police would reduce the number of second jobs they worked and reduce exhaustion (that is, we understood the P → M link) but we were unsure of the degree to which less-exhausted officers would help improve police-community interactions and reduce police use of force against civilians (M → Y).[5] Since union rules in most cities discourage horizontal inequities in treatment of officers, it would be hard to carry out a policy evaluation that randomized officers for pay raises within cities. Instead, any policy evaluation of pay raises for police may well need to be carried out with the police department as the unit of random assignment, which would be an enormously costly way to learn about the overall effects of the policy (M → Y).

Now consider an alternative mechanism experiment: In any big city at any point in time there are officers coming back from a week or two of vacation ("furlough"). Randomly select some police beats but not others within the city to be assigned an officer just returning from furlough, then see what happens to measures of police use of force (or citizen complaints against the police) in those beats that were randomly assigned officers who should be better-rested. This experiment does not test the policy of interest (increased pay), but does tell us something about the causal link we are most unsure of—between the hypothesized mechanism and the outcome of interest—at greatly reduced cost compared to a full-blown policy evaluation. Indeed by demonstrating that manipulation of this candidate mechanism can have causal impacts on the outcome of interest, this mechanism experiment may suggest other policies that operate through that mechanism (reduced officer exhaustion) as candidate policies.

Mechanism experiments can go beyond answering questions about the sign or size of the M → Y link and also illuminate the shape of that relationship. For example, one candidate mechanism through which stepped-up police resources may reduce crime is an increase in the likelihood that victims report to the police (as suggested by Levitt, 1998). However, suppose we did not know whether the effect of additional victim reporting (the M → Y link) gets larger or smaller as the overall level of victim reporting changes; that is, we did not know whether the effect of victim reporting on crime is convex or concave. A very costly way to test this hypothesis is to carry out full-blown policy evaluations that assign increased police presence to some neighborhoods but not others across a large number of neighborhoods. A lower-cost way to test this hypothesis would be a mechanism experiment that randomly assigned rewards for victim reports to the police that result in an arrest in some areas but not in other areas. By exploiting either naturally occurring variation across areas in baseline victim reporting rates, or by

[5] If we were unsure about the effects of the policy (higher salaries) on officers working second jobs, for the sake of our example one could imagine a policy that offered the police union higher salaries in exchange for an agreement that officers would not take on second jobs (the same way that many professional athletes have written into their contracts that they cannot engage in other activities like skiing that can endanger their health).

randomly varying the size of the rewards across areas, we could learn about the functional form for the relationship between M → Y. That would then help us prioritize where to carry out our policy evaluations—that is, where we expect the effect of police on crime to be largest.

3.2 Ruling out policies (and policy evaluations)

Mechanism experiments can also lead to efficiency gains where they can obviate the need for policy experimentation or development by ruling out candidate policies without requiring direct tests of full implementation of those policies. Consider the case where a policy P is under consideration because of a theoretical link to an outcome of interest Y, which is hypothesized to be mediated by mechanism M (or, as in the notation above, P → M → Y). Where the uncertainty is around M → Y, rather than inferring the relationship from analysis of P → Y, we can just test M → Y directly. If M is not linked to Y—and assuming away for the moment any other candidate mechanism by which P could affect Y—we can rule out policies that we hypothesize operate through that M, and move on to more promising avenues of inquiry.

Consider, for example, a policy design question that comes up from time to time related to the Earned Income Tax Credit (EITC), which is whether the EITC would better support goals associated with poverty alleviation and work promotion if it were structured as an earnings supplement, rather than as, in practice, an earnings-linked lump sum transfer. This question is often raised in the context of a larger policy debate around, more generally, whether income supports like the EITC would be better structured as wage subsidies (Phelps, 1994).

Consider a policy, P, for example, to restructure the terms and delivery of the EITC to mimic more closely a wage subsidy, with the goal of improving the welfare, Y, of beneficiaries by advancing the payments and promoting consumption smoothing in a way that we think will increase their utility, dollar for dollar. A policy evaluation that randomly enrolled individuals into either the current EITC or this new policy could answer questions about the welfare impacts of this policy, but would be administratively difficult, expensive, and require overcoming difficult measurement challenges. A mechanism experiment on this same issue sheds important light on the question: researchers encouraged take-up of a little used (and no longer available) option for workers to receive their EITC in earlier and more frequent payments known as the Advance EITC. Promoting this option to beneficiaries, imposing deadline on the choice, and requiring active choice of the way of receiving the credit did not actually increase take-up of the Advance EITC in this sample (Jones, 2010). Taking that revealed preference as an indication that taking up the Advance EITC would not have increased welfare, the experiment provides evidence for policymakers that smoothing payments of the EITC, or consumption out of the EITC, may not be a worthwhile policy change.[6] (In fact, partly based on evidence like this, the Advance EITC was dismantled in 2011.) Recipients do not appear to want to use the EITC in this way; rather, beneficiaries seem to prefer to make use of the lump sum nature of the credit as a form of forced saving.

In addition to ruling out policies, mechanism experiments can also in some cases possibly obviate the need for full-blown policy evaluations. The example from the introduction about an experiment testing the mechanism behind the broken windows theory might provide evidence that would be a basis for forgoing an evaluation of a policy intervention aimed at reducing minor offenses, depending upon the results.

In the case of one of the large-scale social policy experiments from a few decades ago, which studied the national Job Training Partnership Act (JTPA), a federally implemented policy for promoting employment and earnings among dislocated adults and economically disadvantaged adults and youth was evaluated nationally (Bloom et al., 1997). The full program evaluation randomly assigned 21,000 eligible individuals to either receive JTPA services, or not. Under the policy, P, the services provided by JTPA varied by local provider, but generally focused on skill development, and included classroom training, on-the-job training, and other forms of job training. The mechanisms by which the policy was supposed to operate were varied, but very much centered on the idea that the skills and credentials conferred by this type of training—as typified by the general education diploma (GED), receipt of which was in many cases the focus of the training—would allow beneficiaries to command a higher wage. The outcomes of interest were employment and earnings. The evaluation found no positive impact for youth, and only modest positive benefits for adults.

Although it is impossible to know for certain, based on what we now know from other research, this seems like a case where a well-designed mechanism experiment could potentially have at least called into question the need for the full evaluation of a policy such as this. Work by Heckman et al. (2011) finds that the credential of the GED and the type of skills that it reflects are not well correlated with the skills that command wage premia in labor markets, including those paying lower wages. A mechanism experiment could potentially have been designed that examined whether the skills provided through JTPA were valued in the labor market—say in a study where resumes with the sorts of degrees, test scores, and descriptions of skills that would be fostered by JTPA training were sent to employers. If resumes appearing to have JTPA training did not generate greater interest among employers than other resumes, this would have been a signal that a JTPA-style policy evaluation may have been unnecessary.

[6] A new, ongoing experiment with periodic EITC payments (Bellisle and Marzahl, 2015) reaffirms at least the administrative feasibility of an advance EITC.

3.3 Complement other sources of evidence

Experimental evidence has many desirable properties for informing policy, but it is necessarily part of a larger portfolio of policy-relevant evidence. Field experiments exist in the context of other important and useful sources of evidence for informing social policy, including not just policy evaluations but also nonexperimental sources of evidence such as natural or quasi-experiments. Randomized experiments and natural experiments may be complements in a broader program of research on an issue that involves multiple stages (Kling, 2007).

We have argued that one important part of the value of mechanism experiments is to help increase the amount of policy-relevant information that can be obtained for a given research budget, by testing interventions that might not look like an actual (or even feasible) policy. One way mechanism experiments can do that is by increasing the amount of information we can extract from other types of policy-oriented research. Mechanism experiments can help us interpret the results of policy evaluations and quasi-experimental work, including null findings. Once we know that some mechanism is linked to an outcome, the first thing we would check upon seeing a zero impact in a full-scale policy evaluation is whether the policy successfully changed the mediator. Evidence that the mediator was unchanged would suggest the potential value of testing other policies that might generate larger changes in the mediator.

To take an example, consider the conflicting and largely inconclusive body of evidence related to the performance and achievement impacts of school-choice policies (Rouse, 1998; Hoxby, 2003; Cullen et al., 2006; Figlio and Rouse, 2006). Much of this evidence is from quasi-experimental work, although some of it is experimental. The economic theory for the mechanism by which greater choice of schools should lead to improved academic outcomes is fairly straightforward: given greater choices, parents (or whoever is making schooling decisions) can optimize over a choice set of schools, with respect to academic outcomes, and schools can respond to the competitive pressures that are generated. Overall, while the result should be that greater choice leads to improved academic outcomes, the evidence of such effects is scattered. There are a number of points along the causal chain at which this logic could fail to hold, but reduced-form evidence on the effectiveness of school choice does not identify which particular mechanism(s) do not operate as hypothesized.

One mechanism experiment that helps shed light on this mixed evidence was performed by Hastings and Weinstein (2008), who provided actionable, simplified information on school quality to parents. Given that information, parents in the treatment group tended to choose schools with higher test scores. This result unpacks, and provides evidence for, a potential mechanism bottleneck that could explain weak results from other sources of evidence on the effects of school choice. Parents might not have the information necessary, or be able to parse available information effectively, in order to select better performing schools for their children. Moreover, if parents are not doing this effectively, then schools may not be responding to parental choices, either.

Or consider another case, where a policy, P, is intended to affect a particular mechanism, M, under the theory that it mattered for Y. A null finding from evidence looking at the effects of P on Y might occur because P failed to actually affect M, but also might be because M is not linked to Y. A mechanism experiment that shows whether M does or does not matter for Y resolves this.

Even in the case where policy evaluation or quasi-experimental evidence finds that a policy is successful in achieving an outcome of interest, complementary mechanism experiments might still be informative for policy design. New mechanism experiments could also be designed with the explicit goal of better understanding existing natural experiment findings.

For instance, suppose policymakers are concerned that high levels of violence impair the ability of children growing up in distressed urban neighborhoods to succeed in school. This hypothesis is suggested by a series of clever nonexperimental studies that reanalyze population surveys that administer achievement tests to study subjects, and take advantage of the fact that respondents (who hail from different neighborhoods) are assessed at different points in time. Being assessed shortly after a violent event (such as a homicide) occurs in one's neighborhood substantially reduces achievement test scores—on the order of 0.5–0.66 standard deviations in one study (Sharkey, 2010). The size of these effects is enormous, given that for example the black-white test score gap nationwide is typically estimated to be on the order of 0.4–0.8 standard deviations depending on the subject area and age at which tests are administered.[7]

The findings suggest that any policy (P) that reduces violence should improve at least short-term academic outcomes (Y). Unfortunately, this quasi-experimental design is not well suited for telling us about the outcome of primary policy concern—long-term academic outcomes. A policy evaluation that tried to answer this question could become quite costly given the need to administer an intervention substantial enough to reduce crime in distressed neighborhoods and keep it in place for the long term. Such an evaluation would also need some way to deal with the complication of how to measure long-term changes in exposure to violence given residential mobility in and out of the target neighborhoods.

Now consider the following mechanism experiment: One plausible mechanism (M) for the link between exposure to violence and academic outcomes is the effect of crime on stress (Buka et al., 2001). Imagine we identified a sample of people living in

[7] For example the black-white test score gap among 13-year olds in the United States in math is about 0.8 standard deviations in the National Assessment of Educational Progress. On the other hand, the gap measured in the Early Childhood Longitudinal Study of Kindergarteners in reading skills is about 0.4 standard deviations when children start school (Fryer and Levitt, 2004).

high-crime neighborhoods and randomly assigned some to receive long-term enrollment in a meditation-based stress-reduction program (Kabat-Zinn et al., 1992) and then tracked how children did in school over time.[8]

3.4 Understand the role of context in moderating policy effects

A central question facing social policy experimenters is the issue of when and how to export results across contexts. This type of policy forecasting, in which the effects of a policy are estimated before it is put in place, will inevitably require more assumptions, theory, and guesswork than studies on policies that have already been tried (see also Harrison and List, 2004, p. 1033). But policy forecasting is in the end at least as important for public policy. As the distinguished physicist Richard Feynman (1964) once argued, "The moment you make statements about a region of experience that you haven't directly seen, then you must be uncertain. But we always must make statements about the regions that we have not seen, or the whole business is no use." Put differently, in order to forecast the effects of a policy for a new population or in some new geographic context or time period, we need to understand something about the policy's moderators, which can sometimes be facilitated by mechanism experiments that identify mechanisms of actions.

On a practical level, mechanism experiments present a less costly and more practical way to generate direct empirical evidence about the stability of interventions across contexts. Mechanism experiments can be lower-cost ways of understanding how the P → Y link varies across contexts by letting us focus resources on understanding how M → Y link varies across contexts when the M → Y link is the most uncertain link in the causal chain.

Consider for example the US Department of Housing and Urban Development's (HUD's) Moving to Opportunity (MTO) residential-mobility experiment. Since 1994, MTO has enrolled around 4600 low-income public housing families with children and randomly assigned them into three groups: (1) a *traditional voucher group*, which received a standard housing voucher that subsidizes them to live in private-market housing; (2) a *low-poverty voucher group* that received a standard housing voucher that is similar to what was received by the traditional voucher group, with the exception that the voucher could only be redeemed in Census tracts with 1990 poverty rates

[8] Mechanism experiments can also help us build on natural experiment studies by better understanding how to improve policies. That is, if we have a policy that has lots of candidate M's, we could use mechanism experiments to isolate the relative importance of these to design new policies in future that focus more on (and up the dosage for) the key M's. In the previous example, suppose we were unsure about whether exposure to violence mattered because of stress or instead because of, say, depression. We could complement the natural experiment study with two mechanism experiments: one focused on stress (such as meditation) and the other on addressing depression (for example by providing pharmacotherapy). We discuss the possibility of multiple mechanisms in more detail below.

below 10%; and (3) a *control group*, which received no additional services. Assignment to the low-poverty voucher group led to more sizable changes in neighborhood poverty and other neighborhood characteristics than did assignment to the traditional voucher group (Sanbonmatsu et al., 2011). While the traditional voucher arm has the form of a policy experiment, the low-poverty arm had the form of a mechanism experiment—it involved a location restriction that is not a realistic policy option in the US context.

MTO was found to have important effects on both physical and mental health (Kessler et al., 2014; Ludwig et al., 2011a, 2012, 2013; Sanbonmatsu et al., 2011).[9] But the experimental evidence in MTO leaves the precise mechanism generating those effects unidentified, so it is easy to speculate that the causal pathways include mechanisms that either are or are not likely to demonstrate high external validity. So, if those effects happened to operate through, say, something about differences between urban and suburban policing in 1990s in the selected set of cities, we might think the external validity of those results may not be high. If, however, we were able to isolate precisely that MTO effects were due to reductions in experienced stress, that alone improves our ability to make an out of sample forecast because we then more precisely know that what we have to consider is how invariant the relationship is between physical or mental health and stress.

Many “behaviourally informed” policy interventions that are motivated by the view that people are often imperfect decision-makers can be viewed as mechanism experiments that help to elucidate the role of context in policy outcomes. Consider for example the policy question of how the EITC changes individual income, that is, whether the EITC is an effective form of redistribution. One link in the causal chain involves the decision to claim the EITC. Because of the way the policy is implemented in this instance, as a tax credit, the outcome—the degree of income support it provides to recipients—is mediated by claiming and receipt of the credit. Indeed, we observe that there are eligible individuals who fail to receive the credit (and so, any corresponding benefits)—even among individuals who already file income taxes, for whom the marginal costs of claiming the credit are incidental, a portion do not claim the credit (Plueger, 2009). We also observe considerable variation across local areas in how individuals respond to the EITC (Chetty et al., 2013).

One hypothesis about why take-up varies has to do with the potentially moderating effects of variation in information about the credit and knowledge of eligibility on individual use of and response to the credit. In part, because this particular method of income support is administered through the tax code, however, conducting a full-blown policy evaluation experimenting with different versions of the EITC would be impractical for

[9] The same pattern generally holds in the follow-up of MTO outcomes measured 4—7 years after baseline; see Kling et al. (2005), Sanbonmatsu et al. (2006), Kling et al. (2007), and Fortson and Sanbonmatsu (2010).

the purposes about learning about the role of information. But from a mechanism experiment sending timely, simplified notices we see that such an intervention can lead to increased claiming (Bhargava and Manoli, 2013; Manoli and Turner, 2014). This experiment used reminders to test the impact of changes to the mechanism, that is, claiming and receipt of the credit. Simple mailings to a set of roughly 35,000 individuals who appeared eligible for the EITC but did not claim the credit led to significant increases in receipt. However, the effects of these notices faded rapidly and dramatically. In this way, we learn about the extent to which the policy outcomes of the EITC are mediated by claiming frictions generated from the way the credit is currently administered. Policymakers can draw conclusions based on this result for the design of the EITC—most directly, for the way in which the terms of the credit are communicated to eligible individuals, but also potentially for the information requirements those terms reflect.

3.5 Expand the set of policies for which we can forecast effects

Mechanism experiments are not constrained in the same way that policy evaluations are to testing actually feasible or implementable versions of social policies so they can, as a result, test extreme parameter values or unusual functional forms of interventions. Testing unrealistically intensive or pure treatment arms has the benefit of letting us forecast the effects of a wide range of more realistic policy options in those cases when our policy experiments do identify successful interventions. As Hausman and Wise (1985, pp. 194–95) noted 30 years ago: “If, for policy purposes, it is desirable to estimate the effects of possible programs not described by treatments, then interpolations can be made between estimated treatment effects. If the experimental treatments are at the bounds of possible programs, then of course this calculation is easier.”

As a result, while these types of experimentation in social policy can sometimes be viewed as uninformative or irrelevant to policy design, the opposite is the case: by generating information on the nature and range of the behavioral response to an aspect of a policy, mechanism experiments can expand the set of policies for which we can accurately forecast effects. Mechanism experiments can provide a low-cost way to deliver large, even extreme, doses of M to see if the M → Y link matters, and to get a sense of responsiveness of Y to M. By way of comparison, if policy evaluations are constrained to implementable variants of P, and so only manipulate M within the restricted range that allows given the P → M relationship, our understanding of how the policy did or did not work may be inconclusive. If our experiments test interventions that are as intensive as (or even more intensive than) anything that could be accomplished by actual policies, and still don’t work, this lets us rule out policies, as well.

The policy impact that this type of study can have is illustrated by the RAND Health Insurance Experiment that was introduced above (Newhouse and the Insurance

Experiment Group, 1993). Run from 1971 to 1982, this experiment randomly assigned 2750 families to different styles and levels of health insurance coverage. The experiment was designed to provide information on the social welfare impacts of health insurance coverage. The intermediate outcome of interest was behavioral response to health insurance—visits to doctors, hospitals, etc.—and the ultimate outcomes of interest were health outcomes themselves. The central findings were that utilization of health care was responsive to cost sharing, and that overall cost sharing did not have strong effects on health outcomes (though there were some negative effects for lower income participants).

Most notably, the RAND experiment included many treatment arms that do not correspond to any sort of health insurance policy one could buy today. The most generous treatment arm in the RAND experiment offered essentially free coverage, with zero percent coinsurance; other arms were 25%, 50%, and 95% coinsurance rates. Yet this now-decades-old experiment remains one of our most important sources of information about how the generosity of health insurance plans affects the demand for health care and subsequent health outcomes.[10] It continues to be cited heavily even in modern health insurance policy debates, and instrumental to the experiment's prolonged usefulness is the fact that, as a mechanism experiment, it was able to generate such substantial variation in cost-sharing terms in order to observe and estimate behavioral responses.

As another example, in MTO assignment to the low-poverty voucher group led to more sizable changes in neighborhood poverty and other neighborhood characteristics than did assignment to the traditional voucher group (Ludwig et al., 2008). Aside from a few important physical and mental health outcomes, overall, the traditional voucher treatment had relatively few impacts on MTO parents or children through 10–15 years after baseline (Ludwig et al., 2011a, 2012, 2013; Sanbonmatsu et al., 2011). While the low-poverty voucher treatment did not have the sweeping impacts across all outcomes that would be predicted by much of the sociological literature, low-poverty vouchers did generate substantial changes in adult mental and physical health outcomes and overall well-being, had mixed effects on a number of youth outcomes—with girls doing generally better on a number of measures while boys did not. For children who moved to lower poverty neighborhoods when they were relatively young, the treatment led to long-run positive impacts on earnings (Chetty et al., 2016).

Three of us (Congdon, Kling, and Ludwig) have worked on MTO for many years, and have often heard the reaction that the traditional voucher treatment is more policy-relevant and interesting than the low-poverty voucher treatment, because only the

[10] While it was of modest size, it was not cheap. At a total cost of $285 million in 2010, the RAND experiment also holds the record—for now—as the most expensive mechanism experiment ever (Greenberg and Shroder, 2004, p. 181).

former corresponds to a realistic policy option. But it was the low-poverty voucher that generated a sufficiently large “treatment dose” to enable researchers to learn that *something* about neighborhood environments *can* matter for many of these important outcomes, a fact that would not have been discovered if MTO’s design had only included the more realistic traditional voucher treatment. The finding from the low-poverty voucher also provides lessons for why the standard voucher policy has not had such effects (at least in part, it appears, by not inducing sufficient mobility at least in socioeconomic terms). For this reason, findings from the low poverty voucher arm of the experiment have been very influential in housing policy circles.

To take a final example, a policy option sometimes considered to protect workers against a loss of earning power late in their careers is wage-loss insurance (Davidson, 1995; Kletzer and Litan, 2001; LaLonde, 2007). Under most designs of wage-loss insurance, the policy replaces, for covered workers who have lost their job and find reemployment only at a lower wage, some portion of the difference between their older and new wage. The optimal way to set the replacement rate parameter is a question of direct interest for policymakers and researchers. That rate should be set to balance goals of promoting reemployment and supporting consumption while not discouraging search or human capital development. But how individuals will respond is ultimately an empirical question.

In many proposals, the replacement rate is set at 50%; this was also the rate set in a wage insurance demonstration implemented under the *Trade Adjustment Assistance* program. One of the most useful pieces of evidence for informing the design of this policy, however, has been the results of a Canadian experiment with wage-loss insurance that set a replacement rate of 75% (Bloom et al., 1999). That experiment found that covered workers returned to work somewhat faster, but possibly at lower wages. This mechanism experiment testing the functional form of the policy under consideration but with parameter values not under consideration was able to provide information about response patterns that is still useful for policymakers. It is relatively straightforward to interpret the implications of that result for a policy with a 50% replacement, by interpreting the finding as an elasticity. But using relatively extreme values of the policy parameter made it more likely the experiment would precisely estimate a point on the response curve. Even at the larger value of 75%, responses were modest; if a policy experiment had tested a lower replacement rate, the (presumably) relatively smaller responses to the replacement rate would have been harder to detect.

4. WHEN TO DO MECHANISM EXPERIMENTS VERSUS POLICY EVALUATIONS?

The purpose of our paper is not to argue that economists should do only mechanism experiments, or that mechanism experiments are in any sense better than policy evaluations.

Our point instead is that given the relative paucity of mechanism experiments, there may be value in having economists do more of them.

Table 1 presents a framework for thinking about when mechanism experiments can help inform policy decisions. In order for a mechanism experiment to make any sense, we need to believe that we know something about the candidate mechanisms through which a policy might affect outcomes (the key contrast across the columns of Table 1). For a mechanism experiment to be able to tell us something useful about a policy, or to be able to help inform investment of research funding across different candidate policy

Table 1 Policy experiment checklist

	Prior beliefs/understanding of mechanisms	
	Low	**High**
Implications for experimental design	Run a policy evaluation. (or) Do more basic science; use multiple methods to uncover mechanisms.	Run a mechanism experiment to rule out policies (and policy evaluations). (or) Run mechanism experiment to help rule in policies. Either follow with full policy evaluation (depending on costs of policy evaluation, and potential program benefits/scale), or use results of mechanism experiment for calibration and structural estimation for key parameters for benefit—cost calculations.
Implications for policy forecasting/external validity	Run multiple policy evaluations; carry out policy forecasting by matching to estimates derived from similar policies and settings (candidate moderators). Debate: Which characteristics to match on? where do these come from?	Use mechanism knowledge to measure characteristics of policy and setting (moderators) for policy forecasting. Can run new mechanism experiments to test in different settings prior to carrying out policy evaluations in those settings.

evaluations, we either need the list of candidate mechanisms to be "not too long" or to believe that the candidate mechanisms will not interact or work at cross-purposes. Otherwise, information about the causes or consequences of just a subset of mechanisms will be insufficient to either "rule out" any policies, or to identify policies that are worth doing or at least testing and considering further. This contrast is highlighted across the rows of Table 1. The other relevant dimension that varies across the "cells" of Table 1 is the cost or feasibility of carrying out a policy evaluation, which we would always wish to do (regardless of what we had learned from a mechanism experiment) were it cost-less to do so but sometimes is very costly or even impossible.

This framework suggests that under a very particular set of conditions, mechanism experiments by themselves may be sufficient to inform policy decisions. Probably more common are situations in which mechanism experiments and more traditional policy evaluations (which could be either randomized or "natural" experiments) are complements. Under some circumstances, mechanism experiments may not be that helpful and it may be most productive to just go right to running a "black-box" policy evaluation. In this section, we discuss the conditions under which mechanism experiments and policy evaluations will be substitutes and those where they will be complements, and illustrate our key points and the potential scientific and policy impact using different studies that have been carried out.

4.1 Mechanism experiment is sufficient

Mechanism experiments alone may be sufficient to guide policy decisions when economists have some prior beliefs about the candidate mechanisms through which a policy might affect outcomes (and so can design relevant mechanism experiments), while testing the real-world policy lever of ultimate interest is impossible—or at least would entail extraordinarily high cost. Under those conditions, a mechanism experiment could be enough to inform a policy decision if there is just a single mechanism or at least a relatively short list of mechanisms through which the policy may affect outcomes.

If the list of candidate mechanisms through which a policy affects outcomes is "too long" then the only way mechanism experiments could by themselves guide policy would be if we were willing to impose the assumption that the different candidate mechanisms do not have interactive effects. Without this "noninteracting" assumption, a test of one or a subset of candidate mechanisms would not tell us anything of much value for policy since there would always be the possibility that implementing the policy that activated the full menu of mechanisms could have much bigger (or much smaller) effects because of the possibility of interactions among the mechanisms. This condition is likely to be quite rare in practice and so in what follows we focus instead on discussing scenarios under which there is just one mechanism, or there are multiple mechanisms (but not too many of them) that could have interactive effects.

4.1.1 A single mechanism

One scenario under which a mechanism experiment might be enough to guide policy is when there is just a single mechanism (M) that links the candidate policy (P) to the outcome(s) of policy concern (Y). A mechanism experiment is most likely to be sufficient for this purpose if we already understand something about the causal link that carries the policy to the outcome; that is, if we already know either the effect of the policy on the mechanism (P → M), and so just need to learn more about the effects of the mechanism on the outcome (M → Y), or vice versa.

Consider an example from the area of education policy. A key goal of many public policies is to promote college attendance, particularly among low-income people, as a way to achieve redistributional goals and account for positive externalities from schooling attainment. An important open question is the degree to which low levels of college attendance by low-income people is due to the price of college versus the effect of poverty on academic preparation, that is, on how much people learn over their elementary and secondary school careers and so how ready they are to do college-level work. Policies to reduce the price of college among low-income potential college-goers include federal financial aid, especially Pell grants. The existing evidence on the link between financial aid and college attendance has been mixed. Some state-level programs appear to have had large effects, while others have not. In nonexperimental studies, the effects of national changes in the Pell grant program itself have been difficult to disentangle from national changes in other factors affecting college attendance.

This is an example where a mechanism experiment might be enough to guide policy. The candidate mechanism of interest here is price (M), and the key policy question is the degree to which the price of college affects attendance and completion (M → Y). Providing additional financial aid lowers the price (P → M); there are of course questions about the exact magnitude of that relationship and who the "compliers" would be with any given policy change, but at least we can sign that effect. The key puzzle then is to understand the M → Y link. A mechanism experiment can then test this link, while leaving other aspects of the policy environment, such as the "offer" implied by underlying Pell eligibility criteria and the information provided about college that one receives through the Pell application process, as fixed.

The study by Bettinger et al. (2012) builds on the insight that if the key candidate mechanism through which efforts to change educational attainment is the price of college, then potentially *any* intervention that changes this mechanism can provide useful information about the effects of college price on college attendance or persistence (that is, on the M → Y link).[11] Their study generates useful information about the potential

[11] We say "potentially" here because there is a key assumption here about whether the nature of the M → Y link depends on the specific P that is used to modify the value of M; we discuss this issue in greater detail below.

effects of changes to the large-scale Pell grant program by testing an intervention that looks like a change to Pell grant generosity—specifically, the authors worked with H&R Block to increase *take-up* of the existing Pell grants and other federal financial aid programs through the personal assistance of a tax preparer. Note that any other intervention that changed federal financial aid take-up could also have been used. But this particular experiment employed a narrowly targeted form of outreach to customers of tax preparers about whom much financial information was known, and thus probably had much lower costs per additional person aided than broader types of advertising and outreach would.[12]

There are few mechanisms through which the H&R Block intervention might plausibly affect college attendance *besides* receipt of financial aid itself. The most likely alternative mechanism is the possibility of increased general awareness of college and its costs. To examine the empirical importance of this second candidate mechanism, the researchers added a second arm to the experiment which tested the effects of additional general information about college.

The magnitude of the change caused in financial aid received was substantial. For instance, among dependent children whose families received the personal assistance in the experiment, aid increased from $2360 to $3126, on average. This mechanism experiment found that college attendance increased from 28% to 36% among high school seniors whose parents received the personal assistance, and the outcomes of people receiving only additional information were unaffected. We interpret the results as consistent with the idea that the price of college is an important factor determining college attendance, for at least a subset of low-income people; that is, at relatively low cost, we have documented the magnitude of the $M \rightarrow Y$ link. Ideally, we would also do a policy evaluation to better understand take-up rates and the overall magnitude for the change in college price that would result from changing Pell grant generosity, but because the $P \rightarrow M$ link is better understood than the $M \rightarrow Y$ link, the mechanism experiment can be combined with that prior knowledge to generate some additional useful information for policy.

Now consider a different example from the area of urban policy that helps highlight some of the additional assumptions that might be required to rely just on a mechanism experiment to guide policy. A key concern for many cities in the United States is the potential adverse effects on health in high-poverty neighborhoods from the limited availability of grocery stores—so-called "food deserts." The actual policy intervention that is often considered as a response to this potential problem is to subsidize grocery stores to locate into disadvantaged communities. Carrying out a policy evaluation of

[12] Of course, a different policy lever that could be used here is simplification of the process for applying for financial aid, which could potentially also be done at low cost. But a test of this policy change, as with a direct test of changing the Pell grant generosity itself, could only be accomplished through changes in laws.

location incentives for grocery stores would be very costly because the unit of randomization would be the community, the cost per community is high, and the number of communities needed to provide adequate statistical power to detect impacts is large.

The possibility of using a lower-cost mechanism experiment to understand the value of this intervention stems from the plausible assumption that changes in eating healthy foods (fruits, vegetables, whole grains) is the key mechanism (M) through which introducing grocery stores into high-poverty urban areas would improve health, and the recognition that previous research tells us something about the effects of eating healthy foods on health—that is, we already know the M → Y link. Consider the following mechanism experiment that could be carried out instead: Enroll a sample of low-income families, and randomly assign some of them (but not others) to receive free weekly delivery of fresh fruits and vegetables to their homes. By using individuals (rather than communities) as the unit of randomization, this mechanism experiment would be much less expensive than a policy evaluation of the actual policy of interest (subsidized grocery store location). The reduction in costs associated with randomizing people rather than neighborhoods also lets us test a "treatment dose" that is much more intensive than what could be obtained with any realistic policy intervention.

Imagine we found that several hundreds of dollars' worth of free fruits and vegetables delivered to someone's door each month had *no effect* on obesity. This would tell us that even though healthy eating (M) has important impacts on health (Y), changing eating habits (M) through even fairly intensive interventions (P) is challenging in practice. The set of policies about which we could draw conclusions from this mechanism experiment would depend on how much we believe we know about the nature of the P → M link. Suppose we also believed eating habits adapt rapidly to changes in food availability, that social interactions are not very important in shaping eating habits, and that reducing the price of accessing healthy food never *reduces* the chances of eating them (that is, there is a monotonic relationship between the treatment dose and the treatment response); in that case, null results from our mechanism experiment would lead us to predict that *any* sort of policy that tried to address the "food desert" problem would (on its own) be unlikely to diminish problems related to obesity.

If we had more uncertainty about the role of social interactions or time in affecting eating habits, then different mechanism-experiment designs would be required. If we believed that social interactions might be important determinants of people's eating habits, then we would need a more costly experiment with three randomized arms, not just two—a control group, a treatment arm that received free food delivery for themselves, and a treatment arm that received food delivery for themselves and for a limited number of other households that the family designated ("buddy deliveries").[13] If we thought that eating habits were determined at a still larger macro-level, we would have to randomly assign entire communities to receive free home food delivery. A community-level test of home fruit and vegetable delivery could still wind up being less expensive than a policy evaluation of incentive locations for grocery stores, because of the large guarantees that would be required to entice a grocery store to incur the start-up costs of establishing a new location in a neighborhood. But if we thought that eating habits changed very slowly over time, and at the community level, then we would have to commit to providing home food delivery for entire communities for extended periods of time—at which point there might be little cost advantage compared to a policy evaluation of grocery-store subsidies.

4.1.2 Multiple (but not too many) candidate mechanisms

In some situations, it may be possible to learn about the effects of a policy without ever doing a policy evaluation, so long as the list of candidate mechanisms is not "too long." In this case, mechanism experiments can still turn out to be lower-cost ways of generating the necessary policy-relevant information compared to carrying out a full-blown policy evaluation.

Consider a policy (P) that may affect some outcome (Y) through three different candidate mechanisms, given by M_1, M_2, and M_3. If these mechanisms could potentially have interactive effects—that is, the different mechanisms could either amplify or undercut each other's effects—then in a world without resource or feasibility constraints, clearly the best way to test the net effect of the policy would be to carry out a policy evaluation. But sometimes policy evaluations are not feasible, or even if they are, they are enormously costly. In some circumstances, it may be possible to learn about the effect of the policy at lower cost through a mechanism experiment that reduces the cost of learning about at least some of the mechanisms and their interactions with the other mechanisms through interventions that do not look like the policy of interest.

For example, one way to do this is by avoiding the cost of implementing one of the mechanisms (say, M_1) by exploiting naturally occurring population variation in that factor to understand interactivity with the other candidate mechanisms (M_2 and M_3). As an illustration of this idea, consider one of the "kitchen-sink" policy evaluations of the sort that the federal government sometimes supports, like Jobs Plus. This experiment

[13] Duflo and Saez (2003) discuss a cleverly designed experiment that used individuals as the unit of analysis but was designed to identify spillover effects. In their experiment, some people in some departments within a company received incentives to visit a benefit fair to learn more about savings plans. They assessed both direct effects of the information, and effects of information spillovers (from comparisons of the outcomes of the non-incentivized individuals in incentivized departments to individuals in non-incentivized departments). The information diffused through the experiment had a noticeable impact on plan participation.

tested the combined effects of providing public housing residents with financial incentives for work (relief from the "HUD tax" on earnings that comes from setting rent contributions as a fixed share of income—call this M_1), employment and training services (M_2), and efforts to improve "community support for work" (M_3). Previous studies have already examined the effects of the first two program ingredients when administered independently, while the potential value of community support for work is suggested by the work of sociologist William Julius Wilson (1987, 1997) among others. The key program theory of Jobs Plus is that these three mechanisms interact and so have more-than-additive effects on labor market outcomes (Bloom et al., 2005), so carrying out three separate experimental tests of each independent mechanism would obviously not be informative about what would result from the full package. So the bundle was tested with a policy evaluation carried out across six cities, in which entire housing projects were randomly assigned to either a control group or a program group in which residents received the bundle of Jobs Plus services.

What would a lower-cost mechanism experiment look like in this case? Imagine enrolling people who are already living in neighborhoods with high employment rates—so that there is already "community support for work" (M_3) in place "for free" to the researchers. This already makes the intervention being tested look quite different from the actual policy of interest, since the policy is motivated by concern about helping a population that is exactly the opposite of the one we would be targeting—that is, the policy wants to help people in areas with *low* employment rates. Such a design would allow us to capture the interaction of community support for work with other aspects of the policy, although not its main effect. Suppose within these we identify people receiving means-tested housing assistance in those areas, then we randomly assign some of them to receive no reduction in benefits as their income rose (M_1) and employment and training services (M_2).

Our proposed mechanism experiment conserves resources by reducing the dimensionality of the experimental intervention. If we did find some evidence of an effect using this design, we could carry out a follow-up mechanism experiment that included people living in both high- and low-employment neighborhoods—this would let us see how varying the value of M_3 changes the effects of varying the value of the other two mechanisms. This variation in the mechanism is obviously nonexperimental; whether this series of mechanism experiments would dominate just carrying out a full-blown policy evaluation of Jobs Plus would depend partly on how we viewed the trade-off between some additional uncertainty versus additional research costs.

4.2 Mechanism experiment plus policy evaluation

In this section, we discuss different scenarios under which it makes sense to carry out both mechanism experiments and policy evaluations, and provide some examples from previous research. We begin by discussing scenarios in which the mechanism experimentation would come first followed by a policy evaluation, and then scenarios under which the optimal sequence would likely be reversed. Note that even when a mechanism experiment has to be followed by a policy evaluation, the mechanism experiment may still add value by helping us figure out which evaluations are worth running. This includes carrying out mechanism experiments in different settings to determine *where* it is worth trying a policy evaluation.

4.2.1 Mechanism experiments then policy evaluation

Mechanism experiments can help concentrate resources on testing part of a causal chain that links a policy to an outcome. One reason it would make sense to follow a mechanism experiment that had encouraging results with a full-blown policy evaluation would be to learn more about the other parts of the causal chain. An example would be a mechanism experiment that documents that a given mechanism affects some outcome of policy concern ($M \rightarrow Y$), but now for policy purposes we need to also understand the other part of the chain ($P \rightarrow M$). The mechanism experiment can add value here by identifying those applications where the mechanism is unrelated to the outcome ($M \rightarrow Y = 0$) and so avoiding the need to carry out a costly follow-up policy evaluation.

For example, we have argued above that the low-poverty voucher treatment within the MTO residential-mobility demonstration can be thought of as a mechanism experiment—it tests an intervention causing moves to lower poverty areas that is unlikely to ever be implemented as a real policy. This treatment arm makes clear that living in a low-poverty neighborhood of the sort that families with regular housing vouchers move into on their own can have beneficial effects for physical and mental health, delinquency and perhaps even for children's long-term earnings prospects during adulthood. This finding motivates follow-up policy evaluations that test more realistic changes to the voucher policy that might also help steer families into lower poverty areas without an (unrealistic) mandate. Such policies include more intensive mobility counseling or supports compared to what was provided in MTO, or changes in the voucher program design that increases subsidy amounts in lower poverty areas (Collinson and Ganong, 2014).

A different scenario under which it may be worth following a mechanism experiment with a policy evaluation is when implementation of a policy is a significant factor in the causal chain. Medical researchers distinguish between "efficacy trials," which are small-scale research trials of model programs carried out with high fidelity, and "effectiveness trials" that test the effects of some intervention carried out under field conditions at scale. Efficacy trials can be thought of as a type of mechanism experiment. Compared to efficacy trials, effectiveness trials often have more program attrition, weaker training of service providers, weaker implementation monitoring, and smaller impacts

(Lipsey et al., 2007). Thus, an efficacy trial may test the effect of a high-fidelity treatment on a health outcome (M → Y) and an effectiveness trial may show the effect of a policy on provider implementation (P → M) as well as the overall effect of a lower fidelity treatment on health (P → Y).

Sometimes, mechanism experiments can also help highlight lower cost interventions to test with subsequent policy evaluations. Imagine a situation in which we have a policy P_0 attempting to achieve outcome Y, and that P_0 may work through numerous mechanisms $M_1 \ldots M_n$. If a mechanism experiment found a strong effect of M_1 on the outcome, it may be possible to design a simpler policy P_1 that works only through M_1 and is less expensive.

Consider as an example policies that are summer interventions to address the challenges that children from low-income families face maintaining academic gains over the summer. There is a great deal of concern about summer learning loss among poor children relative to more affluent ones. It has long been hypothesized that the loss is due to more limited involvement with academically or cognitively stimulating activities over the summer (Alexander et al., 2007). Potential policy interventions that have been implemented or proposed are to subsidize summer programming for youth (Fifer and Krueger, 2006). To the extent that these interventions look like summer school, they are expensive like summer school.

In this context, we could consider the study by Guryan et al. (2014) as a mechanism experiment that tests one candidate mechanism through which summer school might improve academic outcomes—by increasing the amount of reading students do over the summer. Their study tests this mechanism by sending books directly to the homes of low-income children. The results of that experiment find substantial impacts on reading scores for some students later into the academic year. The implication is that a "summer books" intervention could potentially turn out to be even more cost-effective than summer school itself, and so might warrant a large-scale policy evaluation to calibrate magnitudes.

Since mechanism experiments test an isolated segment of a causal chain, a natural question in this case is to wonder why we do not just test the other parts of the causal chain using separate mechanism experiments. In many cases that might be possible. But one subtle reason this might not work, and so why a follow-up policy evaluation would be required, would be if the link between the mechanism and the outcome (M → Y) depends on the specific policy lever (P) that is used. That is, the (M → Y) link might not be what John DiNardo terms "nonimplementation specific" or what Heckman (2010) calls "policy invariant." In some situations, it might be possible to determine that the (M → Y) link is unlikely to be policy invariant by estimating that relationship in several different mechanisms that manipulate the value of M through some intervention (P) other than the true policy of interest. But in other applications, there may be no substitute for understanding the (M → Y) link when M is manipulated by the actual policy being considered—that is, to do a policy evaluation.[14]

Some simple notation helps illustrate the problem. Let P be the policy, M be the mediator, Y be the outcome (with $P \rightarrow M \rightarrow Y$ as in Fig. 1), with $M = U + V$, $\text{cov}(U, V) = 0$, $\text{cov}(U, Y) = 0$, and $\text{cov}(V, Y) > 0$. That is, only the V part of M is causally related to Y. In population data, we see $\text{cov}(M, Y) > 0$. In this example, M is an implementation specific mediator because policies that change the V part of M will change Y, but policies that change only the U part of M will not influence Y.[15]

4.2.2 Policy evaluation followed by mechanism experiment

The same logic and potential gains come from a mechanism experiment that documents the effects of a policy on some mechanism (P → M), followed by a policy evaluation that helps fill in the effects of the mechanism on the outcome of policy concern (M → Y).

Consider an example from social policy efforts to improve the long-term life outcomes of disadvantaged youth. Recognizing that youth have multiple needs, many interventions in this area bundle together different intervention elements into a single social program. One example of this is the *Becoming a Man* (BAM) intervention, which was designed by Chicago-area nonprofit *Youth Guidance*. BAM is an in-school intervention delivered to youth in groups that uses principals of cognitive behavioral therapy (CBT) to try to get youth to recognize situations in which their automatic responses (what psychologists call "system 1"; see for example Kahneman, 2011) may get them into trouble, and slow down and be more reflective ("system 2") before they act. There are some other candidate mechanisms through which BAM might change youth behavior (such as changes in self-control or "grit") that can largely be ruled out through surveys that the Chicago Public Schools administered to youth in both the treatment and control groups.[16] Yet, one candidate mechanism that many practitioners believe to be quite important for all youth programs is simply putting youth into regular contact with prosocial adults—that is, a basic "mentoring" effect.

A policy evaluation of BAM as implemented in the 2009—10 academic year found the intervention reduces violent-crime arrests by 44% of the mean rate estimated for

[14] One reason we might not see policy invariance is if there is treatment effect heterogeneity in how people's outcomes respond to some mechanism and people also vary in how the value of that mechanism responds to a change in a policy. In this case, who specifically the "compliers" are whose value of M is induced to change by a given P will play an important role in determining what the ultimate effect of the policy is on the outcomes of interest (P → Y). As a real-world example, consider the case of mental health parity for health insurance. Efficacy trials in medicine are able to establish that certain types of mental health treatment improve mental health outcomes. But the effect of the policy on population mental health will depend critically on who the compliers are—the people who whose mental health treatment status changed by the law.

[15] Our thanks to Steve Pischke for this suggestion.

[16] Administrative data rule out the idea that incapacitation is an important effect (since reductions in arrests are not limited to those days when after-school programming is in session).

people who would have participated in the intervention if it had been offered to them (the control complier mean), while a follow-up study in 2013–15 that found similarly large reductions (see Heller et al., 2013; Heller et al., 2015).

Whether the effect of BAM is due to the CBT curriculum itself or instead to a generic "mentoring effect" is of some policy relevance. If the effect were due merely to exposure to a prosocial adult, then any number of nonprofits in Chicago (or anywhere around the country) could be enlisted to provide services to youth, since there would be nothing specific to the BAM curriculum or how it is delivered that would matter. On the other hand, if the content of the CBT curriculum is important, then efforts to deliver that content with fidelity becomes critical. One could imagine following up the BAM policy evaluations with two different types of mechanism experiments. One might have some BAM counselors run versions of the program that essentially threw the curriculum out the window (the weekly group meetings of the youth would be unstructured and just focus on building rapport between the counselors and the youth), or even have youth engage in peer-to-peer mentoring. The other mechanism experiment might (say) enroll youth in the equivalent of an online CBT massive online open course (MOOC) so that there would be no new connection created to any prosocial adult. While these mechanism experiments would have benefits for policy design, it is at the same time not hard to imagine how policymakers (and nonprofit providers) might have had the initial reaction of objecting to the idea of testing what would feel like 'watered-down' versions of BAM that eliminated either the curriculum or the connection to the counselor.

4.3 Just do policy evaluation

A scenario under which the most productive strategy may be to just do a policy evaluation is one in which the policy of interest has a long list of candidate mechanisms that could have interactive effects or work at cross-purposes. Under that type of circumstance, the number of mechanism experiments that would be needed to test different combinations of candidate mechanisms would be large, and because of the possibility of interactive effects it may ultimately require a treatment arm that included all candidate mechanisms. At that point, there is no cost advantage from preceding a policy evaluation with a mechanism experiment—researchers should just go straight to doing a policy evaluation.

Consider for example the effects of changing police staffing levels on crime rates. This is an important policy question because the United States spends over $100 billion per year on police,[17] and hiring more police is an extremely scalable intervention—the one thing that almost every police department in the country can do consistently at large

[17] The figure in 2006 was $99 billion http://www.albany.edu/sourcebook/pdf/t122006.pdf.

scale. Moreover, there remains great debate within the social science community about whether simply putting more police on the street will reduce crime, with most economists of the view that it will while conventional wisdom within criminology remains largely skeptical.

Above, we illustrated the potential value of using mechanism experiments to reduce the costs of understanding treatment effect heterogeneity (by narrowing the set of contexts in which we would need to carry out a policy evaluation) by focusing on a single mechanism through which stepped-up police staffing might affect crime is by changing victim reporting to the police. But in reality, there are many other potential channels as well; for example, police may incapacitate offenders even without victim reporting if police happen upon a crime that occurs in the act. Police presence itself could also directly deter crime, even aside from victims calling the police to report crimes. On the other hand, putting more police on the street could potentially have adverse effects on crime if the result is to exacerbate police-community tensions, or if policing is carried out in a way that reduces perceived legitimacy of the law and the criminal justice system, or if the incapacitation effects of policing are actually negative—that is, if putting more people in jail or prison weakens communities and suppresses informal sources of social control. Understanding the effects of just a subset of these mechanisms would inevitably leave open the key question for policy, which about the net effect of the full bundle of mechanisms that come from putting more police on the street.

The best strategy in this case would be to simply carry out a policy evaluation of what happens from putting more police in some areas but not others. This has been the topic of a large body of work within criminology, in which police departments working with researchers randomly assign extra patrol activity to some high crime "hot spots" but not others; see for example Braga et al. (2012). The one challenge in that literature comes from the possibility of general equilibrium or spillover effects—that is, the possibility that saturating some areas with police could lead criminals to migrate to other areas, or what criminologists call "displacement." In principle, one solution to that problem would be to just carry out random assignment at increasingly large geographic levels. In practice, economists have overcome this problem by relying on natural experiment variation instead (e.g., Evans and Owens, 2007).

A different scenario under which it makes sense to just carry out a policy evaluation directly, without any preceding mechanism experiments, is when the costs of carrying out policy evaluations are very low. This often arises in practice in situations where there is some government service for which there is excess demand, and policymakers use random lotteries as a rationing device. Examples include charter schools or magnet schools, which in many cities and states must use admissions lotteries as a matter of law (see for example Cullen et al., 2006), low-income housing programs, which at present are funded at a level that enables fewer than one-in-four income-eligible households to participate and so leads many cities to use lotteries (see for example Jacob and

Ludwig, 2012; Jacob et al., 2015), and the expansion of Medicaid in Oregon in 2008 (see Taubman et al., 2014; Baicker et al., 2014, 2013; Finkelstein et al., 2012). In our view, randomized lotteries conducted by governments to provide fair access to programs can be turned into field experiments with the appropriate collection of data about all participants in the lottery, regardless of the lottery's outcome.

5. CONCLUSION

In the area of social policy, a great deal of field experimentation is ultimately in the service of informing policy design. If we change the incentives of students and teachers, can we learn how to operate schools to get better educational outcomes? If we vary the structure of health insurance marketplaces, can we learn about how beneficiaries make choices in a way that will allow us to promote broader and cheaper coverage? Questions such as these are at the heart of the movement toward greater use of experimental evidence for social policy design.

The value of a well-executed field experiment is the claim to internal validity—that is, the claim that we have learned something about the effects of the policy of interest in the context in which the policy was tested in the experiment. However, policymakers are often responsible for making decisions about a wide range of contexts beyond those studied in any given policy evaluation. Abstracting from budgetary or feasibility constraints, experimental methods in the form of policy evaluations carried out in different policy-relevant contexts can answer the key questions of policy interest by testing a proposed policy directly. But, in reality, researchers and policymakers alike do in fact face those constraints.

What we have argued in this chapter is that, under some circumstances, the most efficient way to learn about the effectiveness of a policy is not always a direct test of the policy; in fact, what can be most useful are field experiments that bear little surface resemblance at all to the policy of interest. When we have information or beliefs about the *mechanisms* by which policies operate, we can sometimes generate more policy-relevant information per dollar spent by carrying out a mechanism experiment instead of a policy evaluation, and mechanism experiments can sometimes also help improve our forecasts for the contexts under which a policy would be expected to have effects.

Ultimately, then, for researchers and policymakers the issue becomes one of problem selection—what, precisely, should we seek to use field experiments to test? In our view, the portfolio of field experiments in the area of social policy should not consist entirely of mechanism experiments. Policy evaluations will always play a critical role, but there is currently so little attention to mechanism experiments designed to inform policy questions that there may be considerable value in expanding the use of them in practice.

REFERENCES

Alatas, V., Banerjee, A., Hanna, R., Olken, B.A., Purnamasari, R., Wai-Poi, M., 2013. Ordeal Mechanisms in Targeting: Theory and Evidence from a Field Experiment in Indonesia. Working Paper 19127. National Bureau of Economic Research. http://www.nber.org/papers/w19127.

Alexander, K.L., Entwisle, D.R., Olson, L.S., 2007. Lasting consequences of the summer learning gap. Am. Sociol. Rev. 72 (2), 167−180.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics. Princeton University Press, Princeton, NJ.

Angrist, J.D., Pischke, J.-S., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J. Econ. Perspect. 24 (2), 3−30.

Ashraf, N., Berry, J., Shapiro, J.M., 2010. Can higher prices stimulate product use? Evidence from a field experiment in Zambia. Am. Econ. Rev. 100 (5), 2383−2413. http://dx.doi.org/10.1257/aer.100.5.2383.

Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J., Newhouse, J.P., Schneider, E., Wright, B., Zaslavsky, A., Finkelstein, A., The Oregon Health Study Group, 2013. The Oregon experiment − effects of medicaid on clinical outcomes. N. Engl. J. Med. 368 (18), 1713−1722.

Baicker, K., Finkelstein, A., Song, J., Taubman, S., 2014. The impact of medicaid on labor market activity and program participation: evidence from the Oregon health insurance experiment. Am. Econ. Rev. Pap. Proc. 104 (5), 322−328.

Banerjee, A.V., Duflo, E., 2009. The experimental approach to development economics. Annu. Rev. Econ. 1, 151−178.

Bellisle, D., Marzahl, D., September 2015. Restructuring the EITC: A Credit for the Modern Worker. Center for Economic Progress, Chicago, IL.

Bettinger, E.P., Terry Long, B., Oreopoulos, P., Sanbonmatsu, L., 2012. The role of application assistance and information in college decisions: results from the H&R block FAFSA experiment. Q. J. Econ. 127 (3), 1205−1242.

Bhargava, S., Manoli, D., 2013. Why Are Benefits Left on the Table? Assessing the Role of Information, Complexity, and Stigma on Take-up with an IRS Field Experiment. Unpublished Working Paper.

Bloom, H.S., Orr, L.L., Bell, S.H., Cave, G., Doolittle, F., Lin, W., Bos, J.M., 1997. The benefits and costs of JTPA title II-a programs: key findings from the national job training partnership act study. J. Hum. Resour. 32 (3), 549−576.

Bloom, H.S., Riccio, J.A., Verma, N., 2005. Promoting Work in Public Housing: The Effectiveness of Jobs-Plus. MDRC, New York.

Bloom, H., Schwartz, S., Lui-Gurr, S., Lee, S.-W., 1999. Testing a re-employment incentive for displaced workers: the earnings supplement project. Soc. Res. Demonstr. Corp. http://www.srdc.org/media/195754/testing.pdf.

Braga, A., Papachristos, A., Hureau, D., 2012. Hot spot policing effects on crime. Campbell Syst. Rev. 2012, 8.

Buka, S.L., Stichick, T.L., Birdthistle, I., Earls, F.J., 2001. Youth exposure to violence: prevalence, risks, and consequences. Am. J. Orthopsychiatr. 71 (3), 298−310.

Chetty, R., Hendren, N., Katz, L.F., 2016. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. Am. Econ. Rev. 106 (4), 855−902.

Chetty, R., Friedman, J.N., Saez, E., 2013. Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. Am. Econ. Rev. 103 (7), 2683−2721. http://dx.doi.org/10.1257/aer.103.7.2683.

Cohen, J., Dupas, P., 2010. Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. Q. J. Econ. 125 (1), 1−45. http://dx.doi.org/10.1162/qjec.2010.125.1.1.

Cole, S.R., Stuart, E.A., 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. Am. J. Epidemiol. 172 (1), 107−115.

Collinson, R.A., Ganong, P., 2014. The Incidence of Housing Voucher Generosity. Working Paper. http://ssrn.com/abstract=2255799.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Wadsworth.

Cullen, J.B., Jacob, B.A., Levitt, S., 2006. The effect of school choice on participants from randomized lotteries. Econometrica 74 (5), 1191−1230.

Davidson, C., 1995. Wage Subsidies for Dislocated Workers, vol. 95, 31. WE Upjohn Institute for Employment Research.

Deaton, A., 2010. Instruments, randomization, and learning about development. J. Econ. Lit. 48 (2), 424–455.

DiNardo, J., Lee, D.S., 2011. Program evaluation and research designs. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4, Part A. Elsevier, Amsterdam, pp. 463–536.

Duflo, E., Saez, E., 2003. The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment. Q. J. Econ. 118 (3), 815–842.

Evans, W.N., Owens, E., 2007. Cops and crime. J. Public Econ. 91 (1–2), 181–201.

Feynman, R., 1964. A video in the Messenger Lecture Series. Quotation starts at 38:48. The Great Conservation Principles. Available at: http://research.microsoft.com/apps/tools/tuva/index.html#data=4|84edf183-7993-4b5b-9050-7ea34f236045||.

Fifer, M.E., Krueger, A.B., 2006. Summer Opportunity Scholarships: A Proposal to Narrow the Skills Gap. 2006–03. Hamilton Project Discussion Paper. http://www.brook.edu/views/papers/200604hamilton_3.pdf.

Figlio, D.N., Rouse, C.E., 2006. Do accountability and voucher threats improve low-performing schools? J. Public Econ 90 (1), 239–255.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., The Oregon Health Study Group, 2012. The Oregon health insurance experiment. evidence from the first year. Q. J. Econ. 127 (3), 1057–1106.

Fortson, J.G., Sanbonmatsu, L., 2010. Child health and neighborhood conditions: results from a randomized housing voucher experiment. J. Hum. Resour. 45 (4), 840–864.

Fryer Jr, R.G., Levitt, S.D., 2004. Understanding the black-white test score gap in the first two years of school. Rev. Econ. Stat. 86 (2), 447–464.

Greenberg, D., Shroder, M., 2004. The Digest of Social Experiments, third ed. Urban Institute Press, Washington, DC.

Gueron, J.M., Rolston, H., 2013. Fighting for Reliable Evidence. Russell Sage Foundation.

Guryan, J., Kim, J.S., Quinn, D.M., 2014. Does Reading During the Summer Build Reading Skills? Evidence from a Randomized Experiment in 463 Classrooms. Working Paper 20689. National Bureau of Economic Research. http://www.nber.org/papers/w20689.

Hastings, J.S., Weinstein, J.M., 2008. Information, school choice, and academic achievement: evidence from two experiments. Q. J. Econ. 123 (4), 1373–1414.

Harris, J.E., 1985. Macro-experiments versus micro-experiments for health policy. In: Hausman, J., Wise, D. (Eds.), Social Experimentation. University of Chicago Press, Chicago, pp. 145–185.

Harrison, G.W., List, J.A., 2004. Field experiments. J. Econ. Lit. 42 (4), 1009–1055.

Hausman, J.A., Wise, D.A., 1985. Social Experimentation. University of Chicago Press, Chicago.

Heckman, J.J., 2010. Building bridges between structural and program evaluation approaches to evaluating policy. J. Econ. Lit. 48 (2), 356–398.

Heckman, J.J., Humphries, J.E., Mader, N., 2011. The GED. In: Hanushek, E.A., Machin, S., Wo̕βmann, L. (Eds.), Handbook of the Economics of Education,, vol. 3. North Holland, Elsevier, Amsterdam, pp. 423–484 (Chapter 9).

Heller, S.B., Pollack, H.A., Ander, R., Ludwig, J., 2013. Preventing Youth Violence and Dropout: A Randomized Field Experiment. NBER Working Paper 19014.

Heller, S.B., Shah, A.K., Guryan, J., Ludwig, J., Mullainathan, S., Pollack, H.A., 2015. Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. Working paper.

Hotz, V.J., Imbens, G.W., Mortimer, J.H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. J. Econ. 125 (1–2), 241–270.

Hoxby, C.M., 2003. School choice and school productivity: could school choice be a tide that lifts all boats? In: Hoxby, C.M. (Ed.), The Economics of School Choice. University of Chicago Press, pp. 287–342.

Imbens, G.S., 2010. Better late than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). J. Econ. Lit. 48 (2), 399–423.

Jacob, B.A., Ludwig, J., 2012. The effects of housing assistance on labor supply: evidence from a voucher lottery. Am. Econ. Rev. 102 (1), 272–304.

Jacob, B.A., Kapustin, M., Ludwig, J., 2015. The impact of housing assistance on child outcomes: evidence from a randomized housing lottery. Q. J. Econ. 130 (1).

Jones, D., 2010. Information, preferences, and public benefit participation: experimental evidence from the advance EITC and 401(k) savings. Am. Econ. J. Appl. Econ. 2 (2), 147–163.

Kabat-Zinn, J., Massion, A.O., Kristeller, J., Peterson, L.G., Fletcher, K.E., Pbert, L., Lenderking, W.R., Santorelli, S.F., 1992. Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. Am. J. Psychiatr. 149 (7), 936–943.

Kahneman, D., 2011. Thinking, Fast and Slow. Macmillan.

Karlan, D., Zinman, J., 2009. Observing unobservables: identifying information asymmetries with a consumer credit field experiment. Econometrica 77 (6), 1993–2008. http://dx.doi.org/10.3982/ECTA5781.

Keizer, K., Lindenberg, S., Steg, L., 2008. The spreading of disorder. Science 322 (5908), 1681–1685.

Kelling, G.L., Wilson, J.Q., March 1982. Broken windows. Atl. Mon. http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/4465/.

Kessler, R.C., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kling, J.R., Sampson, N.A., Sanbonmatsu, L., Zaslavsky, A.M., Ludwig, J., 2014. Associations of randomization in a housing-mobility experiment with mental disorders among low-income adolescence. J. Am. Med. Assoc. 311 (9), 937–947.

Kletzer, L.G., Litan, R.E., 2001. A Prescription to Relieve Worker Anxiety. Policy Brief 73. Brookings Institution. https://www.piie.com/publications/pb/pb.cfm?ResearchID=70.

Kling, J.R., 2007. Methodological frontiers of public finance field experiments. Natl. Tax J. 60 (1), 109–127.

Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. Econometrica 75 (1), 83–119.

Kling, J.R., Ludwig, J., Katz, L.F., 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing voucher experiment. Q. J. Econ. 120 (1), 87–130.

LaLonde, R.J., 2007. The Case for Wage Insurance. Council Special Report 30. http://www.cfr.org/world/case-wage-insurance/p13661.

Levitt, S.D., 1998. The relationship between crime reporting and police: implications for the use of uniform crime reports. J. Quant. Criminol. 14 (1), 61–81.

Lipsey, M.W., Landenberger, N.A., Wilson, S.J., 2007. Effects of cognitive-behavioral programs for criminal offenders. Campbell Syst. Rev.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. Am. Econ. Rev. Pap. Proc. 103 (3), 226–231.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2012. Neighborhood effects on the long-term well-being of low-income adults. Science 337 (6101), 1505–1510.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L., Kessler, R., Kling, J., Tessler Lindau, S., Whitaker, R., McDade, T., 2011a. Neighborhoods, obesity, and diabetes—a randomized social experiment. N. Engl. J. Med. 365 (16), 1509–1519.

Ludwig, J., Kling, J.R., Mullainathan, S., 2011b. Mechanism experiments and policy evaluations. J. Econ. Perspect. 25 (3), 17–38.

Ludwig, J., Liebman, J., Kling, J., Duncan, G.J., Katz, L.F., Kessler, R.C., Sanbonmatsu, L., 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? Am. J. Sociol. 114 (1), 144–188.

Manoli, D., Turner, N., 2014. Nudges and learning effects from informational interventions: evidence from notifications for low-income taxpayers. Natl. Bur. Econ. Res. Working Papers Series, No. 20718.

Meyer, B.D., 1995. Natural and quasi-experiments in economics. J. Bus. Econ. Stat. 13 (2), 151–161.

Newhouse, J.P., The Insurance Experiment Group, 1993. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press, Cambridge, MA.

Poe-Yamagata, E., Jacob Benus, N.B., Hugh Carrington, M.M., Shen, T., 2011. Impact of the reemployment and eligibility assessment (REA) initiative. IMPAQ.

Phelps, E.S., 1994. Low-wage employment subsidies versus the welfare state. Am. Econ. Rev. 54—58.

Plueger, D., 2009. Earned Income Tax Credit Participation Rate for Tax Year 2005. IRS Research Bulletin.

Rivera, J.A., Sotres-Alvarez, D., Habicht, J.-P., Shamah, T., Villalpando, S., 2004. Impact of the Mexican program for education, health and nutrition (Progresa) on rates of growth and anemia in infants and young children: a randomized effectiveness study. JAMA 291 (21), 2563—2570.

Rouse, C.E., 1998. Private school vouchers and student achievement: an evaluation of the milwaukee parental choice program. Q. J. Econ. 113 (2), 553—602.

Sanbonmatsu, L., Kling, J.R., Duncan, G.J., Brooks-Gunn, J., 2006. Neighborhoods and academic achievement: results from the moving to opportunity experiment. J. Hum. Resour. 41 (4), 649—691.

Sanbonmatsu, L., Ludwig, J., Katz, L.F., Gennetian, L.A., Duncan, G.J., Kessler, R.C., Adam, E., McDade, T.W., Lindau, S.T., 2011. Moving to Opportunity for Fair Housing Demonstration Program—Final Impacts Evaluation. US Department of Housing & Urban Development, PD&R.

Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican Progresa poverty program. J. Dev. Econ. 74 (1), 199—250.

Sharkey, P., 2010. The acute effect of local homicides on children's cognitive performance. Proc. Natl. Acad. Sci. 107 (26), 11733—11738.

Skoufias, E., Parker, S.W., Behrman, J.R., Pessino, C., 2001. Conditional cash transfers and their impact on child work and schooling: evidence from the PROGRESA program in Mexico. Economia 2 (1), 45—96.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., Leaf, P.J., 2011. The use of propensity scores to assess the generalizability of results from randomized trials. J. R. Stat. Soc. Ser. A 174 (2), 369—386.

Taubman, S., Allen, H., Wright, B., Baicker, K., Finkelstein, A., The Oregon Health Insurance Group, 2014. Medicaid increases emergency department use: evidence from Oregon's health insurance experiment. Science 343 (6168), 263—268.

Todd, P.E., Wolpin, K.I., 2008. Ex ante evaluation of social programs. Ann. Econ. Stat. 91/92, 263—292.

US Department of Education, 2015. FY 2016 Department of Education Justifications of Appropriation Estimates to the Congress. http://www2.ed.gov/about/overview/budget/budget16/justifications/index.html.

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.

Wilson, W.J., 1997. When Work Disappears: The World of the New Urban Poor. Vintage.

Wolpin, K.I., 2007. Ex ante policy evaluation, structural estimation, and model selection. Am. Econ. Rev. 97 (2), 48—52.