SOCIAL POLICY:
MECHANISM EXPERIMENTS & POLICY EVALUATIONS

March 28, 2015

William J. Congdon, Jeffrey R. Kling, Jens Ludwig and Sendhil Mullainathan

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, 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 and the American Economic Association 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.

ABSTRACT

Policymakers and researchers are increasingly interested in using experimental methods to inform the design of social policy. In this chapter we develop a framework that helps identify the different types of experimental designs that can be useful for informing policy decisions. We distinguish between policy evaluations, which seek to understand the effectiveness of already designed policies, and mechanism experiments, which do not necessarily examine existing or even feasible policies but instead try to directly test the hypothesized causal mechanisms that are relevant for different policy interventions. We discuss how the selective use of mechanism experiments can complement policy evaluations and increase the amount of policy-relevant information that can be derived for a given research budget, and provide a number of examples from a range of social policy areas including health insurance, education policy, labor market policy, savings and retirement, housing, criminal justice, redistribution, and tax policy. Examples focus on the U.S. context.

William J. Congdon
ideas42
80 Broad Street, 30th Floor
New York, NY 10004
bill@ideas42.org

Jeffrey R. Kling
Congressional Budget Office
2nd & D Streets, SW
Washington, DC 20515
& NBER
Jeffrey_kling@nber.org

Jens Ludwig
Harris School of Public Policy Studies
University of Chicago
1155 East 60th Street
Chicago, IL 60637
& NBER
jludwig@uchicago.edu

Sendhil Mullainathan
Department of Economics
Littauer Center M-18
Harvard University
Cambridge, MA 02138
& NBER
mullain@fas.harvard.edu

I. 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 U.S. Department of Education in 2002 founded the Institute for Education Sciences with a primary focus on running experiments, with an annual budget that was $577 million in 2014 (U.S. Department of Education, 2014). This trend has been spurred in part by numerous independent groups such as the Coalition for Evidence-Based Policy, the Campbell Collaboration, and the Poverty Action Lab at MIT, that promote policy experimentation.

This trend towards 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; 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 results of a 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 non-experimental methods. We argue that an important distinction *between* experimental methods merits greater attention. Specifically in this chapter we argue (and demonstrate through numerous examples) that—perhaps counter-intuitively—the best way to inform social policy is not always to test a policy. Greater use could be made of

[1] Gueron and Rolston (2013) provide an account of this early period in the development of randomized

randomized field experiments that test mechanisms of action through which social policies are hypothesized to affect outcomes, even if the interventions tested do not mimic policies that are actually likely to actually be enacted—what we call *mechanism experiments*.

An example may help to illustrate our argument. Suppose the U.S. 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 basic design was used in a 1960s study by the Stanford psychologist Philip Zimbardo (as described by Kelling and Wilson, 1982, p. 31) and more recently by Keizer, Lindenberg, and Steg (2008), who examined 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; for example, one could hire young men to wear the standard uniform for drug distribution (plain white t-shirt, baggy

jeans, Timberland boots) and have them loiter at randomly selected street corners. Or perhaps a less objectionable version would be to randomly select neighborhoods for 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) the mechanism experiment weakened policymakers' belief in broken windows policing, then we can 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 calibrate magnitudes. 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.

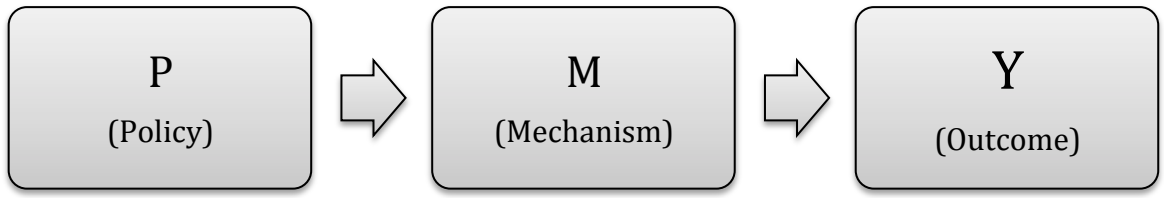
This highlights one important way in which in some cases mechanism experiments can add value—by increasing the amount of policy-relevant information that can be obtained per dollar of research spending. One solution to the concern about external validity with randomized field experiments is replication—that is, testing the policy in many different contexts. As Angrist and Pischke (2010, p. 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."[2] One challenge to this strategy

[2] 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).

stems from resource constraints. Under some circumstances, mechanism experiments can incorporate prior knowledge and focus on the issues about which the most remains to be learned.

In our broken windows example, suppose that from previous work policymakers also know the elasticity of minor offenses with respect to policing (P→M in Figure 1). What policymakers do not know is the accelerator: by how much will reducing minor offenses cascade into reducing other offenses. The mechanism experiment estimates the parameter about which there is the greatest uncertainty or disagreement (M→Y). In contrast, a 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 mechanism experiment concentrates resources on estimating the parameters most relevant to decisions.

Figure 1: Policies, Mechanisms, and Outcomes



The broken windows example is not an isolated case: many policies have theories built into them, even if they are just implicit. Often these theories can be tested more cost effectively and precisely with experiments that do not mimic real (or even feasible) policies. Our view runs counter to the critique some economists leveled against the large-scale government social

experiments of the 1970s and 1980s for "not necessarily test[ing] real policy options" (Harris, 1985, p. 161). 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 play a crucial and under-appreciated role.

With this framework in mind, in the spirit of contributing to a handbook to be of practical use to both policymakers and researchers we organize the remainder of this chapter around two sets of applied questions.

In Section II we answer the question: Why do mechanism experiments? There are two primary motivations. One, as noted above, is to increase the amount of policy-relevant information per research dollar available since replication of policy evaluations is a costly way to learn about external validity. Mechanism experiments can do this by concentrating resources on parameters where policymakers have the most uncertainty, as in the broken windows example, or help rule out policy evaluations that we don't need to run (or for that matter rule out policies), or help us extract more information from other (non-experimental) sources of evidence. A second motivation for carrying out mechanism experiments stems from the recognition that external validity is really a question about 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 III 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.

In both sections we focus on the larger conceptual points behind our argument, but illustrate the potential contributions (and limitations) of mechanism experiments with existing social policy studies whenever possible. For a more comprehensive summary of social policy experiments that have been carried out in the U.S., see Greenberg and Shroder (2004).[3]

II. 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 to that question 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

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

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 (2010, 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 (non-experimental) 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 re-weighting exercise (see, for example, Hotz, Imbens and Mortimer, 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 under-utilized in economics (and the social sciences more generally), we present several

hypothetical examples that illustrate the potential value-added of this approach. 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.

A. 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, that 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→M→Y. If we already believe we understand the P→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, this allow us to learn what we ultimately want to know—about P→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→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.

Consider a hypothetical example, to start: Suppose policymakers are concerned about the secondary consequences of psychosocial stress on poor families, including health impacts such as obesity. For families in poor urban areas, one of the most important sources of stress is crime—particularly gun crime (Cook and Ludwig, 2000; Kling, Ludwig, and Katz, 2005). Policymakers could sponsor a full-scale evaluation of targeted police patrols against illegal guns in high-crime areas, then test the impacts on obesity and other health outcomes. But previous work already tells us something about this intervention's effects on crime (Cohen and Ludwig, 2003), and perhaps also about the effect of crime on stress (Buka, Stichick, Birdthistle, and Earls, 2001). The new information from this experiment is primarily about the stress→obesity link. But for a given budget we could learn more about the stress→obesity pathway (and how that might vary across settings) through a mechanism experiment that enrolled residents of high-crime areas and assigned some to a meditation-based stress-reduction program (Kabat-Zinn et al., 1992).

One important class of parameter uncertainty, on which mechanism experiments can naturally focus, comes about in a situation where for some mechanism M we are confident based on other evidence that P could move M but uncertain whether M will be effective for Y at all. If a mechanism experiment demonstrates that manipulation of a candidate mechanism can have causal impacts on the outcome of interest, this suggests policies that operate through that mechanism as candidate policies. The key cases here are those where existing evidence is thin, or where theory is ambiguous (or silent). Running a full evaluation of a policy P that operates

through that M would be impractical or difficult to justify, but we can learn about the effects of M directly.

Many behaviorally-informed policies, that is, policies that allow for, explicitly address, or make use of the fact individuals are often imperfect decision makers, fall into this class. Behaviorally informed policies often have, explicitly, or even by definition, this property that standard theory does not predict a behavioral response to the mechanisms by which they operate (Mullainathan, Schwartzstein, and Congdon 2012). Such policies are sometimes referred to as "nudges." Because, ex ante, the standard model predicts there is no elasticity to these mechanisms, evidence about the linkage between M and Y is particularly valuable. Running an entire policy evaluation to learn about these effects will typically be impractical. But we can run mechanism experiments to show whether any given Y responds to a particular M, and on that basis say: these are types of lessons that policy might incorporate or things that policy might do. To take one example, consider a set of experiments that tested the impact of notices sent to individuals eligible for the earned income tax credit (EITC). In this case, the policy, P, is the EITC, and the outcome, Y, is the immediate receipt of resources from the credit. The mechanism of interest, M, in this response of individuals to information and knowledge of their eligibility—because this particular method of income support is administered through the tax code, its impacts are mediated by the filing and claiming behavior of eligible individuals. Note that this is a case where individuals eligible for the tax credit should find it their own interest to claim it assuming the amount of the credit outweighs any costs of claiming. Standard theory would not predict a response to manipulation of this mechanism.

In practice, even among eligible individuals who file income taxes, for whom the marginal costs of claiming the credit are incidental, a portion do not claim the credit. An

experiment sending timely, simplified notices does, however, lead to increased claiming (Bhargava and Manoli 2013; Manoli and Turner 2014). In the first experiment, 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. Note that the M identified here as a bottleneck—clear notification to individuals about their eligibility mediates the effect of P on Y—is one that should be under the direct control of policy. That is, this is an M that P could always have addressed—either by addressing understanding of the credit among eligible individuals (or making it less necessary); the M experiment uncovers that it is worthwhile to do so, and suggests the change to policy design, without having to change P directly.

While this type of mechanism experiment often focuses on uncovering informational or attentional mediators of policy effectiveness for outcomes, it need not do so. This type can also uncover behavioral responses to policy parameters that mediate policy impacts despite predictions to the contrary. To take another example drawn from a policy that is administered through the tax code, consider an experiment that varied the presentation of the Savers Credit, a tax credit for retirement savings available to low-income tax filers. The outcome in this case is retirement savings, the policy is the existing credit in the tax code, and the mechanism is the presentation of that credit. An experiment that presented that subsidy as a savings match instead of as a tax credit found that individuals saved more when the benefit was presented as a match (Duflo et al. 2006; Saez 2009). A policy evaluation of the tax credit would have indicated that low-income tax-filers were not very responsive to the subsidy for savings. The mechanism experiment revealed that the presentation of the subsidy was an important mediating factor, and that a presentation which differed from the actual policy led to a considerably larger effect.

Another important source of uncertainty about the M→Y relationship is about its magnitude. To take an example that focuses on that issue, an experiment testing an effort at promoting participation in the banking system at tax filing time scaled demand for a bank account along several dimensions including cost. A policy could offer tax filers a chance to automatically open a bank account into which a tax refund could be deposited and later withdrawn using a bank card providing access to the account. The mediating factors, varied by the mechanism experiment, would be the features and costs of the bank card (Ratcliffe, Congdon, and McKernan 2014). A mechanism experiment focused on determining which features of the card were most attractive could help determine the most promising set of such features, which could then be incorporated into a larger policy involving tax refunds. By independently varying financial and non-financial terms, the financial terms provide a convenient metric that is commonly denominated across contexts and can be used to compare magnitude of the impacts of nonfinancial mediators.

Finally, 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). Suppose for the moment that we believed that this was the only mechanism through which increased police presence might reduce crime, and that we also understood the nature of the P→M link; that is, we knew that having more police on the street reduces the costs to victims of reporting (by reducing wait times) and increases the benefits from a higher chance of apprehending the offender. Assume that in addition we understood the relationship between increased victim reporting (M) and crime (Y), where the M→Y link should (under the standard

economics-of-crime logic) reduce crime through both incapacitation (arresting and imprisoning more criminals) and deterrence (since crimes that police do not know about cannot be punished, increased victim reporting increases the chances of punishment). But suppose that what we did not know is 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 other areas. By exploiting either naturally occurring variation across areas in baseline victim reporting rates, or by 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.

B. Ruling out policies (and policy evaluations)

The famous Iron Law of Evaluation, formulated by Peter H. Rossi (1987, p. 4), considered the discouraging results of the policy evaluation literature of the 1970s and 1980s and stated that: "the expected value of any net impact assessment of any large scale social program is zero."[4] This pessimistic assessment is presumably motivated by the difficulty of consistently implementing social programs well and by our limited understanding of what combination of mechanisms is most important for improving people's life chances. Sometimes mechanism

[4] Rossi's Stainless Steel Law of Evaluation holds that "the better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero." Rossi's Zinc Law of Evaluation is somewhat less pessimistic in its way: "only those programs that are likely to fail are evaluated" (Rossi, 1987).

experiments can be used to provide an initial indication of whether Rossi's law holds in a given context and may do so at reduced cost compared to a series of full-scale policy evaluations.

Under this scenario, mechanism experiments can 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. And 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.

Two examples of mechanism experiments related to the Earned Income Tax Credit (EITC) illustrate the point. The EITC is a refundable tax credit that first increases, and then phases out, with earnings through low levels of earned income. It is intended to increase the returns to working for lower income households, and a now substantial body of research finds positive labor supply responses to the EITC, with little apparent income effect (Meyer and Rosenbaum 2001, Eissa and Liebman 1996). While the EITC was conceived of as an income and consumption support mechanism, in practice it is distributed, and by some evidence used, in a very lumpy way, with claimants taking the credit as part of their tax refunds. A policy design question that comes up from time to time 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. And this is 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 an incremental policy, P, for example, to restructure the terms and delivery of the EITC to mimic more closely a wage subsidy, by, for example changing its default to paying the credit in advance, or automatically moving some of the credit into savings. 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. One experiment that tested a potential mechanism was performed with volunteer income tax assistance (VITA) sites, where individuals were functionally defaulted into saving a portion of their EITC in savings bonds (Bronchetti et al. 2013). Strikingly, given the power of defaults in other contexts, this default had no effect, individuals simply opted out.

More broadly, policy could shift toward paying the EITC at more regular intervals. In fact, the EITC used to offer an option to receive the credit in a smoother fashion, through a little used option known as the Advance EITC. An experiment explicitly focused on testing whether the use of this advance option was mediated by having information about the option to receive the credit this way found that promoting this option to beneficiaries did not have substantial effects on rates of take up (Jones 2010). Providing information about the advance, imposing deadline on the choice, and requiring active choice of the way of receiving the credit had no effect in this sample.

In this case, the value of mechanism experiments has been to provide evidence pointing towards ruling out a policy shift toward smoothing payments of the EITC, or consumption out of the EITC. (In fact, partly based on evidence like this, the Advance EITC was dismantled in 2011.) We learn from the mechanism experiments in this context that policymakers might not

want to change defaults or terms to EITC payments, because the experiments make it clear that recipients do not 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 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 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 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 obviated the need for the full evaluation of a policy such as this. Work by Heckman and others (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 an 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.

C. 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. Margory Turner has put it that "policymaking is a messy, iterative process and the opportunities for evidence to inform and strengthen decisions are numerous and varied. Instead of relying on a single tool policymakers and practitioners should draw from a 'portfolio' of tools to effectively advance evidence based policy" (2013). 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 non-experimental sources of evidence such as natural or quasi-experiments. Policy field experiments and natural experiments in particular 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.

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.

To take an example, consider again the quasi-experimental literature studying the labor supply effects of the EITC (Meyer and Rosenbaum 2001, Eissa and Liebman 1996). The broad findings from that line of research are that the credit has a significant impact on labor force participation, but not on hours—that is, on the extensive margin, but not on the intensive margin. A potential mechanism here involves understanding that there is a subsidy for working additional hours. A mechanism experiment that was in part a test of this was conducted by Chetty and Saez (2013). In this experiment, the researchers worked with a tax preparation firm to randomly provide tax filers who qualified for the EITC with simplified, personalized information about the terms of the EITC, highlighting in particular the way in which the EITC effectively increases the wage of qualifying workers. For example, in the increasing region of the EITC, workers received a message: “Suppose you earn $10 an hour, then you are really making $14 an hour. It pays to

work more!". The experiment finds suggestive evidence that provision of this information leads to increases in earnings in the following year by tax filers receiving this treatment, which provides supporting evidence for the interpretation of the quasi-experimental EITC labor supply research that part of the reason for non-response along the intensive margin is a lack of understanding.

As one additional 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 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. But the evidence of that effect is scattered. It requires that parents (or whomever is making schooling decisions) are themselves optimizing over their choice set of schools, with respect to academic outcomes, and that schools can respond to the competitive pressures that are generated. There are a number of points at which this logic could fail to hold. One mechanism experiment that sheds 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.

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 example suppose we have a policy that has lots of candidate Ms, we could use mechanism experiments to unpack relative importance of these to design new policies in future that focus more on and up the dosage for the key Ms.

For example, numerous studies of compulsory schooling laws document the causal relationship of educational attainment with earnings, crime, health, and other outcomes (Oreopoulos and Salvanes, 2011). Less well understood are the mechanisms behind this relationship. Is it that schooling affects academic skills? Or specific vocational skills? Or social-cognitive skills? The answer is relevant for thinking about how we should deploy the $700 billion the United States spent on K–12 public schooling (data for 2011 from U.S. Department of Education, 2014). Mechanism experiments that assign youth to curricula or supplemental activities that emphasize different specific skills could help policymakers better understand the mechanisms behind the effects of additional schooling attainment.

D. Understand 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 (which non-economists sometimes also call *mediators*).

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 U.S. Department of Housing and Urban Development's (HUD) Moving to Opportunity (MTO) residential-mobility experiment. Since 1994, MTO has enrolled around 4,600 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 below 10 percent; 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 (Ludwig et al., 2008).

MTO was found to have important effects on both physical and mental health (Ludwig et al., 2011, 2012, 2013, Sanbonmatsu et al., 2011).[5] But the experimental evidence in MTO leaves the precise mechanism generating those effects unidentified. And 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.

E. Expand the set of policies for which we can forecast effects

By definition, mechanism experiments are not constrained in the same way that policy evaluations are to testing actually feasible or implementable versions of social policies. Mechanism experiments 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, in spite of Rossi's Iron Law, our policy experiments do identify successful interventions. As Hausman and Wise (1985, pp. 194–95) noted thirty 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."

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

As a result, while these types of experimentation in social policy can sometimes be viewed as uninformative or irrelevant to policy design, but exactly 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 (Newhouse and the Insurance Experiment Group, 1993). Run from 1971 to 1982, this experiment randomly assigned 2,750 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 percent 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.[6] 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., 2011, 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 youth outcomes—with girls doing generally better on a number of measures while boys did not.

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 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

[6] While it was of modest size, it was not cheap. 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).

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. For this reason, findings from the low poverty voucher 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 the their career 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 percent; 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 percent (Bloom et al. 1999). That experiment found that covered workers returned to work somewhat faster, but possibly at lower wages. It is relatively straightforward to interpret the implications of that result for a policy with a 50 percent replacement. But if that experiment had used a lower replacement rate, as an evaluation of a policy with a 50 percent rate would have done, the relatively small responses to the replacement rate mechanism would have been harder to detect.

III. WHEN TO DO MECHANISM EXPERIMENTS VS. POLICY EVALUATIONS?

The purpose of our paper is not to argue that economists should only do 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 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.

Table 1: Policy Experiment Check-List

	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.

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.

A. 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 "non-interacting" 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.

i. 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 re-distributional 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, or instead to the effects of poverty 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 non-experimental 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 credit constraints (M), and the key policy question is the degree to 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.

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).[7] Their study generates useful information about the potential 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.[8]

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.

[7] 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.

[8] 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.

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 percent 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→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 (the P→M link). If that is not feasible, so long as we can sign the effect—believe cov(P,M)>0—then the mechanism experiment alone has at least generated 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 U.S. 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 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").[9] 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.

ii. 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.

---

[9] 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.

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 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, Riccio, and Verma, 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 which 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 non-experimental; 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 tradeoff between some additional uncertainty versus additional research costs.

B. 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.

i. 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→Y), but now for policy purposes we need to also understand the other part of the chain (P→M). The mechanism experiment can add value here by identifying those applications where the mechanism is unrelated to the outcome (M→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 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 there is implementation uncertainty. 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, Landenberger, and Wilson, 2007).

Sometimes mechanism experiments can also help highlight lower-cost interventions to test with subsequent policy evaluations. Imagine a situation in which we have some P or set of P attempting to achieve outcome Y, and operating though a mechanism M that has not yet been experimentally identified. In the case where what we find is that some particular M is doing all of the work—that is, we run an experiment on M, and that seems to explain the entire effect of P

(or is at least as large or larger than the effect of P on Y)—the implication can be that the policy might be just that mechanism.

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 lots of concern about summer learning loss among poor kids relative to rich kids. It has long been hypothesized that the loss is due to more limited involvement with academically or cognitively stimulating activities over the summer (Alexander, Entwisle, and Olson 2007). And 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, Kim and Quinn (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 a separate mechanism experiment. 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 "non-implementation 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.[10]

Some simple notation helps illustrate the problem. Let P be the policy, M be the mediator, Y be the outcome (with P→M→Y as in Figure 1), with M=U+V, cov(U,V)=0, cov(U,Y)=0, and cov(V,Y)>0. That is, only the V part of M is causally related to Y. In population data we see 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.[11]

ii. 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

[10] 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.

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

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 non-profit Youth Guidance. BAM is an in-school intervention delivered to youth in groups that tries to get them 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—a version of what psychologists call cognitive behavioral therapy (CBT). Additional candidate mechanisms through which BAM might change youth outcomes such as dropout and crime involvement include increased exposure to a pro-social adult (and so might induce a sort of "mentoring mechanism"), incapacitation (some youth were able to participate in after-school programming, and so might have avoided criminal activity because they were busy at some after-school activity), field trips to local colleges to help youth learn more about the returns to education, or development of enhanced self-control or emotional intelligence or social skills.

A policy evaluation of BAM as implemented in the 2009–10 academic year found the intervention reduces violent-crime arrests by 44 percent of the mean rate estimated for 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 2013–14 found reductions in overall arrest rates of 29 percent (see Heller, Pollack et al., 2013, and Heller, Shah et al., 2015). One common approach for learning about mechanisms in policy evaluations is to survey the treatment and control group. Surveys that the Chicago public schools collected from youth in the 2009–10 cohort (albeit with a less-than-ideal response rate) include measures of several candidate mechanisms—the small and statistically significant BAM impacts on measures of feeling connected to an adult in the school (relevant to the "mentoring" mechanism), views about the

importance of schooling for the future, and measures of emotional intelligence/social skills and self-control or "grit" seem to suggest those candidate mechanisms may not be very important in practice in explaining BAM's effects on behavior. 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).

To explore the possibility that slowing of automatic responses an important mechanism, Heller, Shah et al. (2015) had youth in the 2013–14 cohort play a modified version of an iterated dictator game, which provoked them to retaliate against unfair behavior by what they believed was another youth at their school (but was a confederate of the research team). They found that BAM increased the amount of time youth spend thinking before they act by nearly 30 percent of the control complier mean. These youth were also randomized to different versions of the game that essentially tried to slow down youth's responses before they decided how much to retaliate against their partners, to prompt all the youth to do what we believe the BAM program gets youth to do on their own: slow down, and reflect on what they are doing before they act.[12]

This follow-up experiment does not test a specific intervention that corresponds to the actual policy of interest (BAM or any actual CBT intervention that would be delivered to youth), but is potentially useful for policy formulation because it helps isolate which mechanism might be most important in driving BAM's impacts on behavioral outcomes. With this information in hand it would be possible to modify the program design to increase the "dosage" of this active ingredient. This information can also be helpful in guiding expansion of the program, since now it is possible to tell providers what key elements of the program they need to have in place and deliver just originally designed versus which ones they can modify to fit local conditions. It

[12] As discussed in detail below one of the randomized conditions in our decision-making experiment does the reverse and tries to "anti-CBT" all youth by getting them to ruminate about the confederate's behavior in the experiment, which yields the same prediction of attenuating the BAM-control difference.

would be natural for some policymakers to question the utility of carrying out a dictator game experiment in the public schools, to have the reaction that this is more about "pure research" than about finding policy levers that work. But this mechanism experiment helped isolate the role of automaticity at much lower cost than could be achieved by carrying out a series of full-scale policy evaluations that vary the mix of program elements that are delivered to youth.

C. 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 U.S. spends over $100 billion per year on police,[13] 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 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.

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

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 assuming that the only 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, Papachristos and Hureau (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, Jacob and Levitt, 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, Kapustin and Ludwig, 2015), and the expansion of Medicaid in Oregon in 2008 (see Taubman et al., 2014, Baicker et al. 2014, 2013, and 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.

IV. CONCLUSION

In the area of social policy, a great deal of field experimentation is ultimately in the service of informing policy design. If we manipulate 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 can be 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? We would never argue that the socially optimal portfolio of field experiments in the area of social policy should 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

Alexander, Karl L., Doris R. Entwisle, and Linda Steffel Olson. 2007. "Lasting Consequences of the Summer Learning Gap." *American Sociological Review* 72 (2): 167–80.

Allcott, Hunt, and Sendhil Mullainathan. 2012. "External Validity and Partner Selection Bias." NBER Working Paper 18373.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." American Economic Review, 80(3): 313–35.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. Mostly Harmless Econometrics. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." Journal of Economic Perspectives, 24(2): 3–30.

Baicker, Katherine, Sarah Taubman, Heidi Allen, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Eric Schneider, Bill Wright, Alan Zaslavsky, Amy Finkelstein, and the Oregon Health Study Group (2013) "The Oregon Experiment – Effects of Medicaid on Clinical Outcomes." *New England Journal of Medicine*. 368(18): 1713-1722.

Baicker, Katherine, Amy Finkelstein, Jae Song, and Sarah Taubman (2014) "The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment." *American Economic Review: Papers & Proceedings*, 104(5): 322-328.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." Annual Review of Economics, vol. 1, pp. 151–78.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment [*]." *Quarterly Journal of Economics* 125 (1): 263–305.

Bertrand, Marianne, and Adair Morse. 2011. "Information Disclosure, Cognitive Biases, and Payday Borrowing." *The Journal of Finance* 66 (6): 1865–93.

Bettinger, Eric P., Bridget Terry Long, Phil Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment." Quarterly Journal of Economics, 127(3): 1205–42.

Bhargava, Saurabh, and Dayanand Manoli. 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*.

Black, Jennifer L., and James Macinko. 2010. "The Changing Distribution and Determinants of Obesity in the Neighborhoods of New York City, 2003–2007." American Journal of Epidemiology, 171(7): 765–75.

Bloom, Howard S., Larry L. Orr. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources* 32 (3): 549–76.

Bloom, Howard S., James A. Riccio, and Nandita Verma. 2005. Promoting Work in Public Housing: The Effectiveness of Jobs-Plus. New York: MDRC.

Bloom, Howard, Saul Schwartz, Susanna Lui-Gurr, and Suk-Won Lee. 1999. *Testing a Re-Employment Incentive for Displaced Workers: The Earnings Supplement Project*. Social Research and Demonstration Corporation. http://www.srdc.org/media/195754/testing.pdf.

Bloom, Howard S., Saskia Levy Thompson, and Rebecca Unterman. 2010. Transforming the High School Experience: How New York City's Small Schools Are Boosting Student Achievement and Graduation Rates. New York: MDRC.

Bronchetti, Erin Todd, Thomas S. Dee, David B. Huffman, and David B. Magenheim. 2013. "When a Nudge Isn't Enough: Defaults and Saving Among Low-Income Tax Filers." *National Tax Journal* 66 (3): 609–34.

Braga, Anthony, Andrew Papachristos, and David Hureau (2012) Hot Spot Policing Effects on Crime. *Campbell Systematic Reviews*. 2012:8.

Buka, Stephen L., Theresa L. Stichick, Isolde Birdthistle, and Felton J. Earls. 2001. "Youth Exposure to Violence: Prevalence, Risks, and Consequences." American Journal of Orthopsychiatry, 71(3): 298–310.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4): 1145–77.

Chetty, Raj, and Emmanuel Saez. 2013. "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients." *American Economic Journal: Applied Economics* 5 (1): 1–31.

Clampet-Lundquist, Susan, Kathryn Edin, Jeffrey R. Kling, and Greg J. Duncan. 2011. "Moving At-Risk Youth Out of High-Risk Neighborhoods: Why Girls Fare Better Than Boys." *American Journal of Sociology*, 116(4): 1154–89.

Cohen, Jacqueline, and Jens Ludwig. 2003. "Policing Crime Guns." In *Evaluating Gun Policy*, ed. Jens Ludwig and Philip J. Cook, 217–50. Washington, DC: Brookings Institution Press.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial." *American Journal of Epidemiology*, 172(1): 107–115.

Collinson, Robert A. and Ganong, Peter (2014). "The Incidence of Housing Voucher Generosity." Working Paper <http://ssrn.com/abstract=2255799>.

Cook, Philip J., and Jens Ludwig. 2000. *Gun Violence: The Real Costs*. New York: Oxford University Press.

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

Cullen, Julie Berry, Brian A. Jacob and Steven Levitt (2006) "The effect of school choice on participants from randomized lotteries." *Econometrica*. 74(5): 1191-1230.

Cutler, David M., Edward L. Glaeser, and Jesse M. Shapiro. 2003. "Why Have Americans Become More Obese?" *Journal of Economic Perspectives*, 17(3): 93–118.

Davidson, Carl. 1995. *Wage Subsidies for Dislocated Workers*. Vol. 95. 31. WE Upjohn Institute for Employment Research.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." Journal of Economic Literature, 48(2): 424–55.

DiNardo, John, and David S. Lee. 2011. "Program Evaluation and Research Designs." In *Handbook of Labor Economics, Volume 4, Part A*; ed. Orley Ashenfelter and David Card, 463-536. Amsterdam: Elsevier.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics,* 118(3): 815–42.

Duflo, Esther, William Gale, Jeffrey Liebman, Peter Orszag, and Emmanuel Saez. 2006. "Saving Incentives for Low- and Middle-Income Families: Evidence from a Field Experiment with H&R Block." *The Quarterly Journal of Economics* 121 (4): 1311–46.

Evans, William N. and Emily Owens (2007) "COPS and Crime." *Journal of Public Economics*. 91(1-2): 181-201.

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

Fifer, Molly E., and Alan B. Krueger. 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.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker and the Oregon Health Study Group (2012) "The Oregon Health Insurance Experiment. Evidence from the First Year." *Quarterly Journal of Economics*. 127(3): 1057-1106.

Fortson, Jane G., and Lisa Sanbonmatsu. 2010. "Child Health and Neighborhood Conditions: Results from a Randomized Housing Voucher Experiment." *Journal of Human Resources*, 45(4): 840–64.

Golomb, Beatrice A., Michael H. Criqui, Halbert White, and Joel E. Dimsdale. 2004. "Conceptual Foundations of the UCSD Statin Study: A Randomized Controlled Trial Assessing the Impact of Statins on Cognition, Behavior, and Biochemistry." *Archives of Internal Medicine*, 164(2): 153–62.

Golomb, Beatrice A., Joel E. Dimsdale, Halbert L. White, Janis B. Ritchie, and Michael H. Criqui. 2008. "Reduction in Blood Pressure with Statins." Archives of Internal Medicine, 168(7): 721–27.

Gotto, Antonio M. 2003. "Safety and Statin Therapy: Reconsidering the Risks and Benefits." Archives of Internal Medicine, 163(6): 657–59.

Greenberg, David, and Mark Shroder. 2004. *The Digest of Social Experiments, 3rd ed.* Washington, DC: Urban Institute Press.

Guryan, Jonathan, James S. Kim, and David M. Quinn. 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.

Gutelius, Margaret F., Arthur D. Kirsh, Sally MacDonald, Marion R. Brooks, and Toby McErlean. 1977. "Controlled Study of Child Health Supervision: Behavioral Results." Pediatrics, 60(3): 294–304.

Harding, David J., Lisa Gennetian, Christopher Winship, Lisa Sanbonmatsu, and Jeffrey R. Kling. 2011. "Unpacking Neighborhood Influences on Education Outcomes: Setting the Stage for Future Research." In W*hither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, ed. Greg J. Duncan and Richard Murnane. New York: Russell Sage Foundation Press.

Hastings, Justine S., and Jeffrey M. Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *The Quarterly Journal of Economics* 123 (4): 1373–1414.

Harris, Jeffrey E. 1985. "Macro-Experiments versus Micro-Experiments for Health Policy." In Social Experimentation, ed. Jerry Hausman and David Wise, 145–85. Chicago: University of Chicago Press.
Harris, Gardiner. 2011. "Federal Research Center Will Help to Develop Medicines." New York Times. January 23, p. A1. Also at: http://www.nytimes.com/2011/01/23/health/policy/23drug.html.
Harrison, Glenn W., and John A. List. 2004. "Field Experiments." Journal of Economic Literature, 42(4): 1009–1055.
Hausman, Jerry A., and David A. Wise. 1985. Social Experimentation. Chicago: University of Chicago Press.
Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In Evaluating Welfare and Training Programs, ed., Charles Manski and Irwin Garfinkel, 201–230. Cambridge, MA: Harvard University Press.
Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." Journal of Economic Literature, 48(2): 356–98.
Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6): 2052–86.
Heller, Sara B., Harold A. Pollack, Roseanna Ander and Jens Ludwig (2013) "Preventing youth violence and dropout: A randomized field experiment." Cambridge, MA: NBER Working Paper 19014.
Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan and Harold A. Pollack (2015) "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago." Working paper.
Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." Journal of Econometrics, 125(1–2): 241–70.
Hoxby, Caroline M. 2003. "School Choice and School Productivity: Could School Choice Be a Tide That Lifts All Boats?" in *The Economics of School Choice*, edited by Caroline M. Hoxby (University of Chicago Press), pp. 287–342.
Imbens, Guido S. 2010. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." Journal of Economic Literature, 48(2): 399–423.
Jacob, Brian A., and Jens Ludwig. 2012. "The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery." *American Economic Review*, 102(1): 272–304.
Jacob, Brian A., Max Kapustin and Jens Ludwig (2015) "The impact of housing assistance on child outcomes: Evidence from a randomized housing lottery." *Quarterly Journal of Economics*. 130(1).
Jepsen, Christopher, and Steven Rivkin. 2009. "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size." Journal of Human Resources, 44(1): 223–50.
Jones, Damon. 2010. "Information, Preferences, and Public Benefit Participation: Experimental Evidence from the Advance EITC and 401(k) Savings." *American Economic Journal: Applied Economics* 2 (2): 147–63.
Kabat-Zinn, J., A. O. Massion, J. Kristeller, L. G. Peterson, K. E. Fletcher, L. Pbert, W. R. Lenderking, and S. F. Santorelli. 1992. "Effectiveness of a Meditation-based Stress

Reduction Program in the Treatment of Anxiety Disorders." American Journal of Psychiatry, 149(7): 936–43.
Keizer, Kees, Siegwart Lindenberg, and Linda Steg. 2008. "The Spreading of Disorder." Science, 322(5908): 1681–85.
Kelling, George L., and James Q. Wilson. 1982. "Broken Windows." The Atlantic Monthly, March. http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/4465/.
Kletzer, Lori G., and Robert E. Litan. 2001. *A Prescription to Relieve Worker Anxiety*. Policy Brief 73. Brookings Institution. https://www.piie.com/publications/pb/pb.cfm?ResearchID=70.
Kling, Jeffrey R. 2006. "Incarceration Length, Employment and Earnings." American Economic Review, 96(3): 863–76.
Kling, Jeffrey R. 2007. "Methodological Frontiers of Public Finance Field Experiments." National Tax Journal, 60(1): 109–127.
Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2005. "Bullets Don't Got No Name: Consequences of Fear in the Ghetto." In *Discovering Successful Pathways in Children's Development: New Methods in the Study of Childhood and Family Life*, ed. Thomas S. Weisner, 243–81. Chicago: University of Chicago Press.
Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.
Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz. 2005. "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment." *Quarterly Journal of Economics*, 120(1): 87–130.
Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics, 114(2): 497–532.
LaLonde, Robert J. 2007. *The Case for Wage Insurance*. Council Special Report 30. http://www.cfr.org/world/case-wage-insurance/p13661.
Liebman, Jeffrey B., and Erzo F. P. Luttmer. 2015. "Would People Behave Differently If They Better Understood Social Security? Evidence from a Field Experiment." *American Economic Journal: Economic Policy* 7 (1): 275–99.
Lazear, Edward. 2001. "Educational Production." *Quarterly Journal of Economics*, 116(3): 777–803.
Leaf, Clifton. 2007. "Why We're Losing the War on Cancer (And How to Win It)." CNN Health, January 9. http://www.cnn.com/2007/HEALTH/01/09/fortune.leaf.waroncancer/.
Levitt, Steven D. (1998) "The relationship between crime reporting and police: Implications for the use of uniform crime reports." *Journal of Quantitative Criminology*. 14(1): 61-81.
Lipsey, Mark W., Nana A. Landenberger, and Sandra J. Wilson. 2007. Effects of Cognitive-Behavioral Programs for Criminal Offenders. Campbell Systematic Reviews.
Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2013. "Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity." *American Economic Review Papers & Proceedings*, 103(3): 226-231.
Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science*, 337(6101): 1505-1510.
Ludwig, Jens, Lisa Sanbonmatsu, Lisa Gennetian, Emma Adam, Greg J. Duncan, Lawrence Katz, Ronald Kessler, Jeffrey Kling, Stacy Tessler Lindau, Robert Whitaker, and Thomas

McDade. 2011. "Neighborhoods, Obesity, and Diabetes—A Randomized Social Experiment." *New England Journal of Medicine*, 365(16): 1509-1519.
Ludwig, Jens, Jeffrey Liebman, Jeffrey Kling, Greg J. Duncan, Lawrence F. Katz, Ronald C. Kessler, and Lisa Sanbonmatsu. 2008. "What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?" *American Journal of Sociology*, 114(1): 144–88.
Malani, Anup. 2006. "Identifying Placebo Effects with Data from Clinical Trials." *Journal of Political Economy*, 114(2): 236–56.
Manoli, Day, and Nick Turner. 2014. "Nudges and Learning Effects from Informational Interventions: Evidence from Notifications for Low-Income Taxpayers." *National Bureau of Economic Research Working Papers Series*, no. 20718.
Mencken, H. L. 1948 [1998]. "Stare Decisis." The New Yorker. Reprinted and abridged in the Wall Street Journal, December 24, 1998, as "A Bum's Christmas." http://www.prismnet.com/gibbonsb/mencken/bumsxmas.html.
Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." Journal of Business and Economic Statistics, 13(2): 151–61.
Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. "A Reduced-Form Approach to Behavioral Public Finance." *Annual Review of Economics* 4.
National Association of Community Health Centers. 2010. "Expanding Health Centers under Health Care Reform: Doubling Patient Capacity and Bringing Down Costs." June. http://www.nachc.com/client/HCR_New_Patients_Final.pdf (on the National Association of Community Health Centers website under "Research $ Data" and "Highlights")
Newhouse, Joseph P., and the Insurance Experiment Group. 1993. Free for All? Lessons from the RAND Health Insurance Experiment. Cambridge, MA: Harvard University Press.
Oreopoulos, Philip, and Kjell G. Salvanes. 2011. "Priceless: The Nonpecuniary Benefits of Schooling." Journal of Economic Perspectives, 25(1): 159–184.
Phelps, Edmund S. 1994. "Low-Wage Employment Subsidies versus the Welfare State." *The American Economic Review*, 54–58.
Ratcliffe, Caroline, William J. Congdon, and Signe-Mary McKernan. 2014. "Prepaid Cards at Tax Time and Beyond." ftp://timecard.urban.org/pubs_prod/2014/pdf/batch1/413082-prepaid-cards-at-tax-time-report.pdf.
Ross, Susan D., I. Elaine Allen, Janet E. Connelly, Bonnie M. Korenblat, M. Eugene Smith, Daren Bishop, and Don Lou. 1999. "Clinical Outcomes in Statin Treatment Trials: A Meta-analysis." Archives of Internal Medicine, 159(15): 1793–1802.
Rossi, Peter H. 1987. "The Iron Law of Evaluation and Other Metallic Rules." Research in Social Problems and Public Policy, vol. 4, pp. 3–20.
Rouse, Cecilia E. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics*, vol. 113, no. 2, pp. 553–602
Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. 2006. "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." Journal of Human Resources, 41(4): 649–91.
Saez, Emmanuel. 2009. "Details Matter: The Impact of Presentation and Information on the Take-up of Financial Incentives for Retirement Saving." *American Economic Journal: Economic Policy* 1 (1): 204–28. doi:10.1257/pol.1.1.204.

Schanzenbach, Diane Whitmore. 2007. "What Have Researchers Learned from Project STAR?" Brookings Papers on Education Policy.

Sen, Bisakha, Stephen Mennemeyer, and Lisa C. Gary. 2011. "The Relationship between Neighborhood Quality and Obesity among Children." In Economic Aspects of Obesity, ed. Michael Grossman and Naci H. Mocan, 145–80. Chicago, IL: University of Chicago Press.

Shah, Anuj K., Sendhil Mullainathan, and Eldar Shafir. 2012. "Some Consequences of Having Too Little." *Science* 338 (6107): 682–85.

Stuart, Elizabeth A., Stephan R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials. The Journal of the Royal Statistical Society, Series A, 174(2): 369–386.

Taubman, Sarah, Heidi Allen, Bill Wright, Katherine Baicker, Amy Finkelstein, and the Oregon Health Insurance Group (2014) "Medicaid Increases Emergency Department Use: Evidence from Oregon's Health Insurance Experiment." *Science*. 343(6168): 263-8.

Thavendiranathan, Paaladinesh, Akshay Bagai, M. Alan Brookhart, and Niteesh K. Choudhry. 2006. "Primary Prevention of Cardiovascular Diseases with Statin Therapy." Archives of Internal Medicine, 166(22): 2307–13.

Todd, Petra E., and Kenneth I. Wolpin. 2008. "Ex ante Evaluation of Social Programs." Annals of Economics and Statistics, 91/92: 263-92.

U.S. Department of Education. 2014. Digest of Education Statistics.

U.S. Department of Education. 2014. "FY 2015 Department of Education Justifications of Appropriation Estimates to the Congress." http://www2.ed.gov/about/overview/budget/budget15/justifications/index.html.

Wehunt, Jennifer. 2009. "The Food Desert." Chicago Magazine, July. http://www.chicagomag.com/Chicago-Magazine/July-2009/The-Food-Desert/.

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

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

Wilt, Timothy J., Hanna E. Bloomfield, Roderick MacDonald, David Nelson, Indulis Rutks, Michael Ho, Gregory Larson, Anthony McCall, Sandra Pineros, and Anne Sales. 2004. "Effectiveness of Statin Therapy in Adults with Coronary Heart Disease." Archives of Internal Medicine, 164(13): 1427–36.

Wolpin, Kenneth I. 2007. "Ex Ante Policy Evaluation, Structural Estimation, and Model Selection." American Economic Review, 97(2): 48–52.

Zhang, Lei, Shuning Zhang, Hong Jiang, Aijun Sun, Yunkai Wang, Yunzeng Zou, Junbo Ge, and Haozhu Chen. 2010. "Effects of Statin Therapy on Inflammatory Markers in Chronic Health Failure: A Meta-Analysis of Randomized Controlled Trials." Archives of Medical Research, 41(6): 464–71.