NBER WORKING PAPER SERIES

MECHANISM EXPERIMENTS AND POLICY EVALUATIONS

Jens Ludwig
Jeffrey R. Kling
Sendhil Mullainathan

Working Paper 17062
http://www.nber.org/papers/w17062

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2011

This paper is forthcoming in the Journal of Economic Perspectives as part of a symposium on field experiments. 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, Tim 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 or National Bureau of Economic Research.

Mechanism Experiments and Policy Evaluations
Jens Ludwig, Jeffrey R. Kling, and Sendhil Mullainathan
NBER Working Paper No. 17062
May 2011
JEL No. C93

ABSTRACT

Randomized controlled trials are increasingly used to evaluate policies. How can we make these experiments as useful as possible for policy purposes? We argue greater use should be made of experiments that identify behavioral mechanisms that are central to clearly specified policy questions, what we call "mechanism experiments." These types of experiments can be of great policy value even if the intervention that is tested (or its setting) does not correspond exactly to any realistic policy option.

Jens Ludwig
University of Chicago
1155 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

Jeffrey R. Kling
Congressional Budget Office
and NBER

Sendhil Mullainathan
Department of Economics
Littauer 208
Harvard University
Cambridge, MA 02138
and NBER
mullain@fas.harvard.edu

I. INTRODUCTION

Randomized controlled trials are increasingly used to evaluate policies. To cite one example, in 2002, the U.S. Department of Education founded the Institute of Education Sciences with a primary focus on running experiments; its annual budget is now about $700 million (U.S. Department of Education, 2010). This trend has been spurred in part by numerous independent groups—the Coalition for Evidence-Based Policy, the Campbell Collaboration, an international network of researchers hosted by the Norwegian Knowledge Centre for the Health Services, the Poverty Action Lab, and Innovations for Poverty Action—that promote policy experimentation. Others however question the wisdom of this trend. A vigorous debate has arisen around 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). We argue this debate has often been framed too narrowly on experimental versus non-experimental methods. An important distinction *between* experimental methods has been overlooked.

Suppose a policy maker has already decided on using an experiment. She faces a design problem. Given a fixed budget, how should she design her experiment to maximize policy-relevant information? The answer seems obvious: replicate the policy as it would be implemented at scale, and randomly assign units (people or sites of the sort that would be targeted by the policy) to treatment and control conditions. The design challenges involve selecting the most cost effective units of randomization and the data collection strategies. We call the resulting experiments *policy evaluations.* In practice most policy experimentation involves policy evaluations. Yet in some (practically

relevant) situations, these are not the best experiments to use—even if the sole goal is to help inform policy decisions.

A simple example illustrates our point. 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 credibly exogenous source of variation in the implementation or intensity of broken windows policing across areas, which rules out the opportunity for a low-cost study of an existing "natural experiment" (Meyer, 1995, Angrist and Pischke, 2009). To an experimentally-minded economist, the most obvious next step goes something like: DOJ should choose a representative sample of cities, randomly select half of their high-crime areas to receive broken windows policing (or perhaps randomly assign half the cities to get citywide broken windows policing), and carry out a traditional *policy evaluation*.

Now consider an alternative experiment: Buy a small fleet of used cars. Break the windows of half of them. Park the cars in a randomly selected subset of neighborhoods, and then measure whether more serious crimes increase in response. What might seem like a fanciful example is actually the basic research design used in the 1960s study by Stanford psychologist Philip Zimbardo that helped motivate the broken windows theory (Kelling and Wilson, 1982, p. 31),[1] which in turn led to the implementation of broken windows policing at scale in New York City during the 1990s. One can of course perform

[1] The same design was used more recently by a Dutch team for a study published in *Science* (Keizer et al., 2008).

variants with other small crimes; for example, one could hire young men to wear the standard-issue uniform for drug distribution (plain white t-shirt, baggy jeans, Timberland boots) and have them loiter at randomly selected street corners. This *mechanism experiment* does not test a policy: it directly tests the causal mechanism that underlies the broken windows policy.

Which experiment would be more useful for public policy? Partly it's an issue 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 policy makers' 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 very strong effects, we might now run a policy evaluation to calibrate magnitudes. Or, 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.

Mechanism experiments more carefully incorporate prior knowledge and can be designed to maximize information in the places where the policy maker needs to know the most. In our broken windows example, suppose there is general agreement about the list of minor offenses that might plausibly accelerate more serious crimes (that is, the list of candidate mediating mechanisms M in Figure 1). Suppose (from previous work) we 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 Figure 1). 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 city / year combination. With enough sample (that is, money), one could recover the (M→Y) link. In a world of limited resources, mechanism experiments concentrate resources on estimating the parameters that are most decision relevant.

We argue that mechanism experiments should play a more central role in the policy process. The broken windows example is not an isolated case: many policies have theories built into them, even if they are sometimes 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 argument runs counter to the critique leveled by some economists against the large-scale government social experiments of the 1970s and 1980s for "not necessarily test[ing] real policy options" (Harris, 1985, p. 161). We argue that some of these experiments, because they highlight mechanisms, could have far-reaching *policy* value. Social scientists already value mechanism experiments because they contribute to building knowledge. Our argument is that *even if the sole goal were informing policy*, mechanism experiments play a crucial, under-appreciated role.

This distinction between mechanism experiments and "policy evaluations could also change the debate between about the use of experimentation to guide policy. We feel many of the criticisms of experimentation are really criticisms of policy evaluations—particularly "black box" policy evaluations where the mechanisms through which the

policy may affect outcomes are numerous or unclear. Deaton (2010, p. 246) for example fears experimentation generates information that is too "narrow and local" to be of much use for policy. While this can also be true for mechanism experiments, because of their emphasis on *how* programs work, the knowledge gained can extend to a broader range of situations.

The next section of the paper provides a brief review of how economists have thought about experimentation and the problem of forecasting the effects of different types of policies in different settings – that is, the challenge of external validity. We then discuss what mechanism experiments can teach us. We use the randomized Moving to Opportunity (MTO) residential mobility experiment (and several hypothetical extensions) as an extended example. We then suggest a framework to help think about the conditions under which the most policy-relevant information comes from a mechanism experiment, a policy evaluation, or both, and close with some suggestions for future research.

II. POLICY EXPERIMENTS AND EXTERNAL VALIDITY

Policymaking is inevitably about prediction. What is the effect of some familiar policy when implemented in the future, or in some new setting? What is the effect of some entirely new policy? A useful way to think about the types of research activities that help answer these questions comes from Wolpin (2007) and Todd and Wolpin (2008). They distinguish 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 DiNardo and Lee (2010, p. 2) describe as beginning "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 if it 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."

The challenge in making policy forecasts from *ex post* evaluations stems from the possibility that treatments may interact with characteristics of the policy's setting – including the target population, time period, or other contextual factors. The effects of broken windows policing in Evanston, an affluent North Shore suburb of Chicago, may differ from the policy's effects when implemented in distressed neighborhoods on the south side of Chicago. Those features of a policy's setting (or of the policy itself) that may influence the policy's impacts are what the research literature outside of economics calls *moderators*. We argue that the type of experiment that is most useful for addressing

this challenge and informing policy forecasts depends on what researchers believe they know about a policy's mechanisms, which non-economists sometimes also call *mediators*.

At one extreme are situations in which researchers do not know very much about a policy's candidate mechanisms -- the list of plausible mechanisms might be overwhelmingly long, or we might have little sense for whether the mechanisms potentially interact or even work at cross purposes, or what (if any) aspects of the relevant causal chain operate in ways that are invariant across policy settings. The standard approach has been to carry out policy evaluations in as many settings as possible of the sort in which the policy might actually be implemented. 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). 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."

As mentioned above, there is an active debate within the economics profession about the value of this type of research program, focused largely on issues of external validity. Under this approach we forecast a policy's effects in some setting using previous tests of the policy in similar settings – that is, we try to match on the policy's candidate moderators (see, for example, Hotz, Imbens and Mortimer, 2005, Cole and Stuart, 2010, Imbens, 2010, Stuart et al., 2011). One challenge is that without some understanding of a

policy's mechanisms, how do we decide which aspects of the policy or its setting is a potentially important moderator? Another challenge comes from the fact that policy evaluations are costly, and so we will never be able to carry out evaluations of ever candidate policy of interest in every potentially relevant setting. We have nothing new to add to this debate, which we view as largely orthogonal to the main argument we advance in this paper.

The type of situation to which our paper is relevant arises when researchers have some beliefs about the mechanisms through which a policy influences social welfare. One way researchers currently use such beliefs is by interpreting the results of randomized experiments through the lens of a particular structural model (Wolpin, 2007; Todd and Wolpin, 2008; Heckman, 2010; Imbens, 2010). This approach takes the policy experiment that is run as given, imposes some assumptions after the fact about the policy's mechanisms, fits the model, then forecasts the effects of a wide range of policies and settings. This approach can make sense when we have sufficiently sharp prior beliefs about the way the world works to be confident that we have the right structural model, and that the key structural parameters really are structural (that is, invariant across settings). The structural model substitutes assumptions for data, traded off against the risk that our assumptions are incorrect.

But if we believe we know something about the mechanisms through which the policy might operate, why limit ourselves to using this information only after a policy evaluation has been designed and carried out? Why not use this information to help inform the design of the experiment that is being run? Why not design experiments that are explicitly focused on isolating the effects of candidate mechanisms? Once our focus

shifts to identifying mechanisms, the importance of having close (or even any) correspondence between the interventions we test and the specific policy applications we seek to inform is diminished. The change that this way of thinking implies for the design of our policy experiments is not just cosmetic. The change can be drastic, as we illustrate in the next section.[2]

Mechanism experiments can help with the policy forecasting or external validity problem in two ways. First, improved understanding of a policy's mechanisms can help us predict what aspects of the policy or its setting may moderate the policy's impacts. Second, by taking advantage of what researchers believe they already know mechanism experiments can be less costly than policy evaluations. This means that we can carry out relatively more mechanism experiments in different settings, and help prioritize the types of policies and settings in which we should carry out full-scale policy evaluations.

III. MECHANISM EXPERIMENTS

In what follows we illustrate some of the ways in which mechanism experiments can help generate policy-relevant information. We use the Moving to Opportunity (MTO) residential-mobility experiment as an extended example. As context for this example, we note that economists have become increasingly interested in the role of social interactions in affecting people's choices and behavioral outcomes (Becker and Murphy, 2000; Manski, 2000). Housing policy affects the social interactions that people experience (as well as the quality of their local public goods) by affecting the geographic concentration of poverty in America.

[2] Another relevant observation here is that if we really believe that the key structural parameters in our model are structural, then there is no intrinsic reason that we would need to test a real policy to identify their value.

To learn about the effects of concentrated neighborhood poverty on poor families, in the early 1990s the U.S. Department of Housing and Urban Development launched the MTO demonstration. 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). The traditional voucher treatment did not have many detectable impacts on the outcomes of MTO parents or children 4-7 years after baseline (Kling, Ludwig and Katz, 2005, Sanbonmatsu et al., 2006, and Kling, Liebman and Katz, 2007, Fortson and Sanbonmatsu, 2010). The low-poverty voucher treatment generated a more complicated pattern of impacts. The low-poverty voucher did not affect children's schooling outcomes, perhaps because MTO moves wound up generating only modest changes in school quality, but did cause sizable reductions in youth violence. The low-poverty voucher had effects on other youth behaviors that differed by gender, with girls doing better and boys doing worse as a result of the moves (Clampet-Lundquist et al., 2011). For adults, the low-poverty voucher treatment did not product detectable changes in labor market or other economic outcomes, but did have

important effects on mental health and some physical health outcomes, including obesity.[3]

Two of us (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 *can* matter for important outcomes like adult obesity, 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 illustrate the different ways in which mechanism experiments might be useful for policy, we focus on one of the key findings from MTO —the reduction in adult obesity. Imagine New York City policymakers were interested in reducing the prevalence of obesity, which contributes to health problems like diabetes and heart disease. One pattern in the epidemiological data that would immediately be obvious to policy planners would be the large disparities in obesity prevalence across neighborhoods. Data from 2003-7 show that the share of adults who are obese ranges from 8 percent on the Upper East Side to 30 percent in the directly adjacent neighborhood of East Harlem (Black and Macinko, 2010).

These patterns would presumably lead policymakers to think about the "production function" that produces obesity, in the spirit of Sen et al. (2009). Of course

[3] The public health community officially defines obesity as having a body mass index, which is weight in kilograms divided by height in meters squared, or BMI = kg/m^2, of 30 or more.

the prevalence of obesity is a function of calories consumed and time spent in energetic activity, but also of other factors that interact with diet and physical activity, and that vary across neighborhoods. For example, some neighborhoods may be "food deserts" that have few grocery stores that sell fresh fruits and vegetables, which might lead people to eat higher-calorie, less healthy foods (Wehunt, 2009). Some neighborhoods may have few parks or places to exercise. Neighborhood conditions might also affect levels of psychosocial stress, which in turn can affect diet and exercise.

Depending on our prior beliefs about certain aspects of the framework discussed above, mechanism experiments could be useful for guiding policy in this area by helping us: 1) rule out candidate policies; 2) expand the set of policy options for which we can forecast effects; 3) prioritize available research funding; 4) concentrate resources on estimating parameters about which we have the most uncertainty or disagreement; and 5) strengthen causal inference with either randomized or "natural" experiments.

1. Ruling out policies

Twenty-five years ago the distinguished sociologist Peter H. Rossi (1987, p. 4) considered the discouraging results of the policy-evaluation literature of the 1970s and 1980s and formulated his Iron Law of Evaluation: "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 about what combination of mechanisms is most

[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."

important for improving people's life chances. Sometimes mechanism experiments can be used to provide an initial indication of whether Rossi's law holds in a given context, and to do so at reduced cost compared to carrying out a series of full-scale policy evaluations.

For example, policymakers concerned that distressed neighborhoods may be "food deserts" might consider a policy of subsidizing grocery stores to locate into such areas as a way to reduce obesity. Carrying out a policy evaluation of location incentives for grocery stores would be expensive because the unit of randomization is the community, the cost per community is high, and the number of communities needed to have adequate statistical power is large.

Now consider the following mechanism experiment that could be carried out instead: Enroll a sample of low-income families, and randomly assign some of them to receive free weekly delivery of fresh fruits and vegetables to their homes. By using individuals as the unit of randomization, rather than communities, this mechanism experiment would be much less expensive than the more "realistic" policy evaluation. 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. 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 fruits and vegetables 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 the obesity problem.

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").[5] 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.

[5] 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.

The possibility of using a mechanism experiment to learn more about "food deserts" is not an isolated example. Consider that the 2010 health care legislation included $9.5 billion in funding over five years to double the number of people served by local community health centers. Expanding community health care centers could potentially reduce total health care spending by improving access to regular preventive health care services, thereby reducing the need for emergency care (National Association of Community Health Centers. 2010). However, this hypothesis has not been subject to any sort of rigorous empirical test that we know of, because it is hard to imagine running a large-scale experiment that randomly assigned areas to get community health care centers. But we could run a mechanism experiment that randomly assigned individual low-income families in medically under-served neighborhoods to receive improved access to routine, preventive health care using a van that pulled right up in front of their home. Indeed, this was the same basic idea behind the Washington, D.C. Mobile Unit for Child Health Care experiment that was run 45 years ago (Gutelius et al., 1977).

As a final (non-health) example, consider the question of whether smaller high schools improve student achievement. Possible mechanisms through which this might occur include stronger relationships between students and school staff, having students spend more time around peers who share their interests, and providing school administrators with more autonomy (Bloom et al., 2010). Instead of immediately carrying out a full-scale, very costly policy evaluation of small schools, why not first carry out a mechanism experiment focused on bonding instead? Take a representative sample of charter schools, which already provide administrators with autonomy. Randomly assign some teachers to be offered the chance to earn overtime pay by working after school and

weekends with small, randomly-selected groups of students in an effort to promote faculty-student and student-to-student bonding. Evidence that this intervention was capable of promoting student engagement and academic outcomes would suggest the value of carrying out a large-scale policy evaluation. But evidence that even what H.L. Mencken (1948) would call a "horse-doctor's dose" of extra bonding did not affect student outcomes would greatly reduce the motivation to carry out a large-scale policy evaluation of smaller schools.

2. *Expand the set of policies and settings for which we can forecast policy impacts*.

Ruling out entire classes of policy interventions is easier when our experiments test interventions that are as intensive (or more) as anything that could be accomplished by actual policies. Testing unrealistically intensive treatment arms also 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, p. 194-5) noted a quarter-century 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."

The policy impact that this type of study can have is illustrated by the RAND Health Insurance Experiment, which included many treatment arms that do not correspond to any sort of health insurance policy one could buy today (Newhouse et al., 1993). Yet this 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. With a total cost of $285 million in 2010 dollars, the RAND experiment also holds the record – for now – as the most expensive mechanism experiment of all time (Greenberg and Shroder, 2004, p. 181).

3. Prioritize research funding

If a mechanism experiment tested the most intensive imaginable intervention to address the problem of "food deserts" and found no effect on obesity, we would rule out not only the value of policies to address food deserts but also, obviously, the value of policy evaluations to test those types of policies. Null results from a policy evaluation of a more realistic but less-intensive intervention would not let us shut down an entire line of research inquiry in the same way, since it would always be possible to imagine that a slightly more intensive intervention might yield more promising results.

Encouraging results from a mechanism experiment would help us decide where to invest additional research funding, and might also help shape the types of policies that we subjected to full-scale policy evaluations. Suppose, for example, we found that delivering hundreds of dollars worth of free fruit and vegetables to the doorsteps of low-income families each month only changed the consumption of, say, apples. This finding might lead policymakers to conclude that the right policy to evaluate is not just a costly effort to incentivize new grocery stores to move into high-poverty areas, but also (or perhaps instead) a lower-cost program to subsidize bodegas and convenience stores just enough to make it profitable for them to stock this one type of fruit.

4. Concentrate resources on estimating parameters about which we are most uncertain.

In the introduction we noted that mechanism experiments can help us concentrate resources on estimating parameters about which we have the most uncertainty or disagreement. As another example along these lines, suppose policymakers are concerned about the secondary consequences of psychosocial stress on poor families, including health impacts. For families in poor urban areas, one of the most important sources of stress is crime – particularly gun crime (Cook and Ludwig, 2000; Kling, Liebman and Katz, 2005; 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 et al., 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) by carrying out 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).

In other situations we might be most uncertain about the link between our policy levers and key mediating mechanisms (P→M). For example, in the case of obesity we already understand the connection of body mass with diet and exercise (Cutler, Glaeser, and Shapiro, 2003). But we might not understand the effects of our policies on diet and exercise. In situations like this, diet and exercise become what medical researchers call "surrogate clinical endpoints," which then become the dependent variables of interest for

our experiments. The idea of focusing selectively on testing individual links in a causal chain also raises the possibility of using mechanism experiments to compress the timetable required to learn about the long-term effects of some policy, by testing different sequential links in a causal chain simultaneously.

Perhaps less obvious is the value of carrying out multiple experiments that use different policy levers to manipulate the same mechanism, given the great difficulty of determining what is the true mediating mechanism that links a policy to an outcome, rather than just a proxy for the mediating variable that really matters.[6] Showing that the effects of reduced stress on obesity is the same regardless of whether stress levels are modified through a meditation program or by some sort of anti-gun policing program would be informative about whether the mediating mechanism of stress is "non-implementation specific," to use John DiNardo's term, or what Heckman (2010) calls "policy invariant."

A final non-health example about the ability of mechanism experiments to focus research resources comes from the possibility of avoiding the need to carry out full-blown "synergy" (or "kitchen sink") experiments of the sort that the federal government regularly sponsors, 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), employment and training services, and efforts to improve "community support for work." Previous studies have already examined the effects of the first two program

[6] Some simple notation suggested to us by Steve Pischke 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.

ingredients when administered independently, while the potential value of community support for work is suggested by the Wilson (1987, 1996) 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). Across six cities, Jobs Plus randomly assigned entire housing projects to either a control group, or a program group in which residents received the bundle of Jobs Plus services.

We could have instead carried out a mechanism experiment that enrolled a slightly less disadvantaged (and hence slightly less directly policy-relevant) study sample that needed one or two but not all three of the mechanisms the Jobs Plus theory suggests are needed for labor market success. Imagine enrolling people who applied for means-tested housing assistance, which in some cities is rationed using randomized lotteries (Jacob and Ludwig, 2011), and are already living in neighborhoods with high employment rates. Then we randomly assign some of them to receive employment and training services. A test of the Jobs Plus "synergy" theory comes from comparing the response to these services for those who were versus were not lucky enough to be randomly assigned a housing subsidy. Our proposed mechanism experiment conserves resources by reducing the dimensionality of the experimental intervention.

5. Help strengthen causal inference

Mechanism experiments can help us interpret the results of policy evaluations, 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. Without the mechanism experiment, we wouldn't be sure whether it would be worth following up a null impact from a policy evaluation with more research in that area.

Mechanism experiments can also strengthen the basis for causal inference with "natural experiment" policy evaluations. Imagine a simple pre-post study of aggregate U.S.-level data of a change in Medicaid policies that reduced out-of-pocket costs to poor adults from having, say, bariatric surgery, in which part of someone's stomach is removed or reduced in size by a gastric band in order to reduce appetite and food consumption. Suppose the study found that after the policy change, bariatric surgery rates among low-income people increase and obesity rates decline. Absent any additional information, this study design would not provide compelling evidence about the link between bariatric surgery and obesity, given the large number of other factors that are changing over time that influence obesity. But there would seem to be far fewer confounding threats to estimating the effect of the policy on bariatric surgery rates (the P→Y link) from a simple pre-post comparison. Additional evidence about the mechanism (the M→Y link between bariatric surgery and obesity) would enable us to infer how much of the time trend in obesity prevalence was due to the Medicaid policy change. Evidence for the mechanism-outcome link here happens to come from a medical trial rather than an experimental test of an unrealistic policy, but the example nonetheless highlights our key point.

New mechanism experiments could even be designed with the explicit goal of better understanding existing natural experiment findings. 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, 2009). 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 $485 billion the U.S. spends each year on K-12 public schooling (U.S. Census Bureau. 2011, Table 258). Why not spend a few million dollars on a mechanism experiment that assigns youth to curricula or supplemental activities that emphasize different specific skills, to better understand the mechanisms behind the effects of compulsory schooling laws?

IV. WHEN CAN MECHANISM EXPERIMENTS BE USEFUL?

The purpose of our paper is *not* to argue that economists should *only* carry out mechanism experiments, or that mechanism experiments are "better" than policy evaluations. Our main point is that given the current paucity of mechanism experiments designed to help answer policy questions, on the margin we think that economists should be doing more of them.

Table 1 presents a framework for thinking about the conditions under which mechanism experiments can help inform policy decisions. Under a very particular set of conditions, mechanism experiments may by themselves be sufficient to guide policy decisions. More common are likely to be scenarios in which mechanism experiments and traditional policy evaluations (which could include "natural" as well as randomized experiments) are complementary. Under some circumstances mechanism experiments

might not even be that helpful, and a more useful approach would be to just go right to running a black-box policy evaluation.

1. *When Mechanism Experiments Can Be Helpful*

In order for a mechanism experiment to make any sense at all, we need to believe that we know at least something about the candidate mechanisms through which a policy might affect the outcomes of ultimate policy concern (the right-hand column of Table 1).

Under some circumstances mechanism experiments might be sufficient to guide policy design. We need to believe that the list of candidate mechanisms through which a policy might affect outcomes is fairly short, or that the long list of potentially relevant mechanisms do not interact or work at cross purposes (a short list of candidate mechanisms that could interact would not by itself preclude a mechanism experiment). Depending on the application we might need to know something already about other parts of the causal chain. At the very least we would need to be confident that existing systems are capable of reliably delivering the policies that activate key mechanisms. Even then, if the cost of carrying out a policy evaluation were low enough relative to the policy stakes, we would probably still wish to carry out a policy evaluation to improve our policy forecast. We would settle for just a mechanism experiment if the costs of carrying out a policy evaluation were prohibitive, or the policy stakes were low.

2. *Do Mechanism Experiments plus Policy Evaluations*

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

other parts of the causal chain, such as when there is implementation uncertainty. For example, 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, since having a bespectacled, laptop-toting professor loom over the program's implementation is not usually standard operating procedure. Compared to efficacy trials, larger-scale effectiveness trials often have more program attrition, weaker training for service providers, weaker implementation monitoring, and smaller impacts (Lipsey et al., 2007).

As noted above, prior evidence from mechanism experiments can enhance the efficiency of our portfolio of policy evaluations by helping us figure out which evaluations are worth running. This includes carrying out mechanism experiments in different settings to determine in where it is worth trying a policy evaluation.

Learning about the mechanisms through which a policy affects outcomes can also help predict which aspects of the policy or its settings will moderate the policy's impacts, although it is worth keeping in mind that the correspondence between mechanisms and moderators is far from perfect. For example, the well-known Tennessee STAR experiment found that reducing class sizes in elementary school improved learning outcomes (Krueger, 1999; Schanzenbach, 2007). Lazear (2001) argues that a key mechanism for these class-size effects is the reduced chance that instructional time in a given classroom is diverted by disruptive students, which helps explain why in the STAR experiment lower-income and minority students seemed to benefit the most. But when California had to hire a large number of teachers to enact class-size reduction statewide

average teacher quality seemed to decline, particularly in those schools serving disproportionately low-income and minority students (Jepsen and Rivkin, 2009). Thus, teacher quality turned out to be a surprise mechanism (at least to California policymakers). Student background wound up being a moderator that influenced different parts of the causal chain in different ways.

3. Do Some Combination of "Basic Science" and Policy Evaluation

In some situations researchers do not yet know enough to narrow down the list of candidate mechanisms through which a policy operates, or worry that a policy's long list of candidate mechanisms might interact (or if some might work at cross purposes) – represented by the first column of Table 1. The debate within the economics profession is about whether it is best under these circumstances to carry out "basic science" studies or to carry out policy evaluations. The extreme position is that policy evaluations can *never* be useful for policy purposes, which strikes us as unlikely to be correct.

In the case of Moving to Opportunity, for example, observational studies going back to the 1920s had shown that neighborhood attributes are correlated with behavioral outcomes, even after controlling for individual- and family-level factors. Policymakers need to know whether such correlations reflect an underlying causal relationship, which is relevant to decisions like whether to whether to devote resources to building public housing or to private-market rent subsidies, or whether to allow suburban townships to limit zoning approval for low-cost housing. Given the large number of potentially-interacting mechanisms through which residential location might affect behavior and

well-being, it is not clear that anything short of a black-box policy evaluation would have much value in guiding these policy decisions.

One common criticism of black-box policy evaluations is that we cannot understand the characteristics that explain heterogeneity of treatment effects (that is, a policy's moderators) without understanding the policy's key mediating mechanisms. While there is no question that evidence about mechanisms is tremendously valuable, we believe it is not correct that black-box evaluations are never useful.

Consider the example of statins, which have been used since the late 1980s[7] and shown in numerous randomized clinical trials to reduce the risk of heart disease and overall mortality (Ross et al., 1999, Gotto, 2003). Statins were originally thought to prevent heart attacks by lowering cholesterol levels in the blood, which in turn reduced the chance of plaque build-up. But meta-analyses of black-box clinical trials showed that statins improved health outcomes even among people who already had relatively low levels of blood cholesterol at baseline (Golomb et al., 2004, Wilt et al., 2004, Thavendiranathan et al., 2004). Moreover, these meta-analyses showed that the cardiovascular benefits of statins seemed to occur too rapidly after onset of treatment to be explained by the effects of statins on plaque accumulation (Golomb et al., 2008). The leading hypothesis – at least for now – is that statins reduce heart attacks partly by reducing inflammation (Zhang, 2010) or blood pressure (Golomb et al., 2008).

The key point for present purposes is that right now we don't really know exactly why statins reduce heart attacks. Yet meta-analyses of black-box clinical trial studies show that they clearly do improve health, and can also tell us something about how their effects on health are moderated by patient characteristics such as age, gender, and

[7] Thanks to Elbert Huang and Harold Pollack for this example.

baseline health status. Our limited understanding of the mechanisms through which statins work has not prevented them from becoming one of the world's top-selling drug classes, to the extent that some medical experts have suggested should be "put into the water supply" (Golomb et al., 2004, p. 154).

A similar point was made during Congressional testimony in 1971 by Sidney Farber, the "godfather of cancer research," who argued (as quoted in Fortune, 2007): "We cannot wait for full understanding; the 325,000 patients with cancer who are going to die this year cannot wait; nor it is necessary, to make great progress in the cure of cancer, for us to have the full solution of all the problems of basic research… The history of medicine is replete with examples of cures obtained years, decades, and even centuries before the mechanism of action was understood for these cures – from vaccination, to digitalis, to aspirin."[8]

This is not to say that later understanding of mechanisms does not generate tremendous benefits to society. For example, learning more about how chemotherapy works has dramatically increased the benefit/cost ratio of such treatments over time. But evidence that an intervention works, even if we don't understand why, is better than not having access to that intervention at all. Repeated black-box experiments can eventually help us learn something about the policy's moderators, and, as in our statins example, can also inform our theorizing about candidate mechanisms as well.

[8] Thanks to Harold Pollack for suggesting this quotation. At the risk of over-emphasizing the point, one more example comes from two of the most important mental health drug discoveries – lithium, which is used to treat bipolar disorder, and Thorazine, which is used to treat psychosis. Modern medicine has very little understanding of why either medicine works in helping patients (Harris, 2011).

V. CONCLUSIONS

It seems like common sense that the best way to use experiments to inform policy is to test policies. However, we argue here for increased use of randomized experiments that identify behavioral mechanisms that are central to clearly specified policy questions, even if the specific interventions that are tested (or their settings) do not correspond exactly to what policymakers would implement in practice. While our suggestion might seem obvious once articulated, mechanism experiments that are designed to help answer specific policy questions remain rare. We hasten to add that mechanism experiments and traditional policy evaluations are as a general proposition best thought of as complements, rather than substitutes. We need to make greater use of mechanism experiments, on the margin, without fetishizing mechanisms.

The larger question of how to structure experiments to maximize the ability to apply the findings more generally in other contexts opens up a number of potentially fruitful lines of additional research beyond what we have considered here. For example, many people seem to have the intuition that evidence about either the link between policy levers and mechanisms ($P \rightarrow M$ from Figure 1) or between mechanisms and outcomes ($M \rightarrow Y$) is more generalizable than evidence about the link between policies and ultimate outcomes of interest ($P \rightarrow Y$). It is not hard to think of situations in which this is true, but this need not be true in all cases.[9] It would be useful to learn more about how the causal

[9] Imagine a case with a single candidate mediator and outcome of interest. Whether either of the individual links in this causal chain ($P \rightarrow M$ or $M \rightarrow Y$) is more stable across contexts than is the total effect of the policy on the outcome, $P \rightarrow Y$, depends in part on how $P \rightarrow M$ and $M \rightarrow Y$ co-vary across contexts. It is not hard to imagine cases in which the two relationships negatively co-vary, so the effect of the policy on the outcome is more stable across situations than the link between the policy and mediator or the mediator and outcome. Suppose that in neighborhoods where adoption of broken windows policing leads to relatively larger increases in arrests for minor offenses, the stigma of arrest declines, and so the deterrent effect of the prospect of being arrested goes down. Or suppose that in areas where local residents are not very easily

links from $P \rightarrow M$ and $M \rightarrow Y$ co-vary across contexts, and the extent to which those links reinforce each other or may tend to offset each other.

A second line of investigation that seems worth exploring more is the benefits and costs of policy field experiments (both mechanism experiments and policy evaluations) versus "natural experiment" studies. Sometimes natural experiment studies have designs that are as good as random assignment of treatment because they actually involve random assignment (see for example, Angrist 1990, Kling, 2006, and Jacob and Ludwig, 2011). But more often, natural experiment studies necessarily rely on research designs that generate information that may be more local than that obtained from an experiment (such as regression discontinuity), or that may be more vulnerable to omitted variables bias. On the other hand, natural experiment studies circumvent the external validity concerns raised by either randomization bias (the self-selection of people willing to sign up for a randomized experiment; see Heckman, 1992 and Malani, 2006) or selection-partner bias (the willingness of organizations to participate in experiments; see Alcott and Mullainathan, 2011). But with few exceptions little is currently known about the extent of randomization or selection-partner bias in practice. Alternatively, policy field experiments and natural experiments may be complements in a broader program of research on an issue that involves multiple stages (Kling 2007).

A final question worth considering is the issue of when and how to export results across contexts. While statistical matching of estimates obtained from similar interventions and contexts is fine, as far as it goes, a broader framework would let us incorporate behavioral models, parameters, and prior beliefs into the policy forecasting

---

deterred by the prospect of being arrested, policymakers respond by implementing this policing strategy in a way that leads to relatively larger numbers of minor arrests.

exercise. This type of policy forecasting, or *ex ante* policy evaluation, will inevitably require more assumptions, theory and guesswork than *ex post* studies of previous policies (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 haven't seen, or it's no use in the whole business."

REFERENCES

Allcott, Hunt and Sendhil Mullainathan (2010) "External validity and partner selection bias." Working Paper, Harvard University Department of Economics.

Angrist, Joshua D. (1990) "Lifetime earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*. 80: 313-335.

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.

Banerjee, Abhijit V. and Esther Duflo (2009) "The experimental approach to development economics." *Annual Review of Economics*. 1: 151-178.

Beatty, Alexandra (2009) *Strengthening Benefit-Cost Analysis for Early Childhood Interventions: Workshop Summary*. Washington, DC: National Academy of Sciences Press.

Becker, Gary S. and Kevin M. Murphy (2003) *Social Economics: Market Behavior in a Social Environment*. Belknap Press of Harvard University Press.

Black, Jennifer L. and James Macinko (2010) "The changing distribution and determinants of obesity in the neighborhoods of New York City, 2003-7." *American Journal of Epidemiology*.

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 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.

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.

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): 1154-1189.

Cohen, Jacqueline and Jens Ludwig (2003) "Policing crime guns." In *Evaluating Gun Policy*, Jens Ludwig and Philip J. Cook, Eds. Washington, DC: Brookings Institution Press. pp. 217-250.

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.

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.

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.

Deaton, Angus (2010) "Instruments, randomization, and learning about development." *Journal of Economic Literature*. 48: 424-455.

DiNardo, John and David S. Lee (2010) "Program evaluation and research designs." Cambridge, MA: NBER Working Paper 16016.

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): 815-842.

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

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): 840-864.

Fortune, Clifton Leaf. 2007. "Why we're losing the war on cancer (and how to win it)." CNN Health. January 9. At <http://articles.cnn.com/2007-01-09/health/fortune.leaf.waroncancer_1_gamma-rays-testicular-cancer-national-cancer-act/6?_s=PM:HEALTH>.

Golomb, Beatrice A., Michael H. Criqui, Halbert White, and Joel E. Dimsdale (2004) "Conceptual foundations of the UCSD statin study." *Archives of Internal Medicine*. 164: 153-162.

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): 721-727.

Gotto, Antonio M. (2003) "Safety and statin therapy." *Archives of Internal Medicine*. 163: 657-659.

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

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: 294-304.

Harris, Jeffrey E. (1985) "Macro-experiments versus micro-experiments for health policy." In Jerry Hausman and David Wise, Eds. *Social Experimentation*. Chicago: University of Chicago Press. pp. 145-185.

Harris, Gardiner. 2011. "Federal Research Center Will Help to Develop Vaccines." *New York Times*. January 23, p. A1. 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.

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

Hastings, Justine. S., and Lydia Tejeda-Ashton (2008) "Financial Literacy, Information, and Demand Elasticity: Survey and Experimental Evidence." Cambridge, MA: NBER Working Paper No. 14538.

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*, Edited by Charles Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press. pp. 201-230.

Heckman, James J. (2010) "Building bridges between structural and program evaluation approaches to evaluating policy." *Journal of Economic Literature*. 48(2): 356-398.

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: 241-270.

Imbens, Guido S. (2010) "Better LATE than nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*. XLVIII(2): 399-423.

Jacob, Brian A. and Jens Ludwig (2011) "The effects of housing assistance on labor supply: Evidence from a voucher lottery." *American Economic Review*.

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): 223-250.

Kabat-Zinn, J., AO Massion, J Kristeller, LG Peterson, KE Fletcher, L Pbert, WR Lenderking and SF Santorelli (1992) "Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders." *American Journal of Psychiatry*. 149: 936-943.

Keizer, Kees, Siegwart Lindenberg, and Linda Steg (2008) "The spreading of disorder." *Science*. 322: 1681-1685.

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/

Kling, Jeffrey R. (2006) "Incarceration length, employment and earnings." *American Economic Review*. 96(3): 863-876.

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*, Edited by Thomas S. Weisner. Chicago: University of Chicago Press. pp. 243-281.

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.

Lazear, Edward (2001) "Educational production." *Quarterly Journal of Economics*. 116(3): 777-803.

Lipsey, Mark W., Nana A. Landenberger, and Sandra J. Wilson (2007) *Effects of Cognitive-Behavioral Programs for Criminal Offenders*. Campbell Systematic Reviews.

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*.

Manski, Charles F. (2000) "Economic analysis of social interaction." *Journal of Economic Perspectives*. 14(3): 115-136.

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." At <http://www.io.com/gibbonsb/mencken/bumsxmas.html>.

Meyer, Bruce D. (1995) "Natural and quasi-experiments in economics." *Journal of Business and Economic Statistics*. 13(2): 151-161.

National Association of Community Health Centers. 2010. "Expanding Health Centers Under Health Care Reform: Doubling Patient Capacity And Bringing Down Costs." June. *At* <http://www.nachc.com/client/HCR_New_Patients_Final.pdf>.

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 (2009) "How large are returns to schooling? Hint: Money isn't everything." Cambridge, MA: National Bureau of Economic Research Working Paper 15339.

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: 1793-1802.

Rossi, Peter H. (1987) "The iron law of evaluation and other metallic rules." *Research in Social Problems and Public Policy*. 4: 3-20.

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): 649-691.

Schanzenbach, Diane Whitmore (2007) "What have researchers learned from Project STAR?" *Brookings Papers on Education Policy*.

Sen, Bisakha, Stephen Mennemeyer, and Lisa C. Gary (2009) "The relationship between neighborhood quality and obesity among children." Cambridge, MA: NBER Working Paper 14985.

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): 369-286.

Thavendiranathan, Paaladinesh, Akshay Bagai, M. Alan Brookhart, and Niteesh K. Choudhry (2006) "Primary prevention of cardiovascular diseases with statin therapy." *Archives of Internal Medicine*. 1666: 2307-2313.

Todd, Petra E. and Kenneth I. Wolpin (2008) "Ex ante evaluation of social programs." Working Paper, University of Pennsylvania Department of Economics.

U.S. Census Bureau. 2011. *Statistical Abstract of the United States.* At <http://www.census.gov/compendia/statab/>.

U.S. Department of Education. 2010. "Fiscal Year 2011 Budget Summary — February 1, 2010. Section III. F. Institute of Education Sciences." At <http://www2.ed.gov/about/overview/budget/budget11/summary/edlite-section3f.html>.

Wehunt, Jennifer. 2009. "The Food Desert." *Chicago Magazine.* July. At <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*. 464: 1427-1436.

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: 464-471.

Figure 1 – Logic Model for Broken Windows Policing

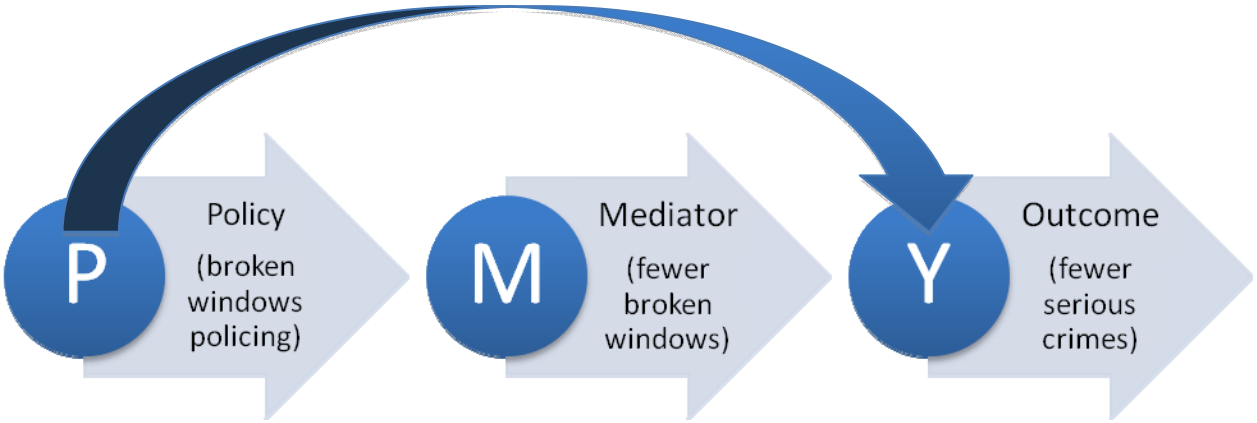

Table 1
Policy Experiment Check-List

	Prior beliefs / understanding of mechanisms	
	Low	**High**
Implications for experimental design	Run a policy evaluation OR Do more basic science; 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.