Evaluating Public Health Law Using Randomized Experiments

A Methods Monograph

PHLR

Making the Case for Laws that Improve Health

PUBLIC HEALTH LAW RESEARCH

January 25, 2012

Robert Wood Johnson Foundation

Evaluating Public Health Law Using Randomized Experiments

A Methods Monograph
for the
Public Health Law Research Program (PHLR)
Temple University Beasley School of Law

By:

Alan S. Gerber, Ph.D.
Department of Political Science
Yale University

Donald P. Green, Ph.D.
Department of Political Science
Columbia University

Allison Sovey, M.A.
Yale University

PHLR is a national program of the Robert Wood Johnson Foundation

Summary

The randomized clinical trial (RCT) research design has transformed medical research and is now accepted as the most reliable method for measuring the effects of drugs and other specific medical interventions. In this monograph we describe how randomized trials can be used to evaluate broader issues related to public health laws and policies. The distance between a medical study and evaluating a public policy appears vast. The unit of observation in medical studies is almost always the individual patient, and interventions are usually simple, clearly defined treatments, such as alternative drugs or protocols. In contrast, to evaluate laws, policies, or programs experimentally involves randomly exposing people or communities to multifaceted interventions that are often implemented on a grand scale. To study the effects of a poverty relief program on health, for example, scholars randomly raise the incomes of selected households or regions. To assess effects of alternative political processes for selecting public works projects, some regions are randomly assigned one method of voting and other regions randomly assigned an alternative.

The appeal of random assignment is clear in principle. A core question is whether it is feasible for investigating the effects of laws, policies and programs. Although random assignment of laws to communities is understandably rare (see Forster et al., 1998 for a more noteworthy exception), one can readily imagine how implementation of policies and programs might lend itself to random allocation. In this monograph we will show that, in fact, the use of randomized experiments to study public policy and institutional design not only can be done, but is in fact already underway and gaining momentum. We begin by introducing a basic system of notation for defining and analyzing treatment effects, known as the potential outcomes model (Rubin, 1990; Splawa-Neyman, 1923). We then review basic properties of the RCT and the assumptions that are required for randomized experiments to produce unbiased estimates of treatment effects.

Randomization alone does not ensure that a comparison of experimental groups will yield estimates of useful quantities. Assumptions are required for group comparisons to yield estimates of the treatment effect of interest, and further assumptions are required for generalization from the particularities of a given trial; an especially important issue when experiments are used to forecast a policy's effect in a different setting. Typically, RCTs must be conducted on a much smaller scale than the program they are intended to evaluate, and so a critical issue is whether an RCT can accurately foretell effects of a scaled up version of the intervention. After discussing assumptions that are required for the use of experiments to measure effects of interest, we then describe some recent applications of the RCT to questions of public policy and institutional design. A brief review of the recent literature shows a range of how randomized experiments are used on issues related to health and social policies. We conclude with reflections on the direction of recent research and prospects for future work.

This monograph is restricted to a discussion of field experiments and naturally occurring real-world randomizations, that is, interventions conducted in naturalistic settings in which the treatments and outcomes are those relevant in real world contexts. Social science laboratory experiments (Webster & Sell, 2007) can illuminate specific micro-level mechanisms of behavioral effects. However, the question of whether results obtained in laboratory settings apply in real-world settings remains an open one, and so we focus on field experiments, where the gap between the experimental setting and the policy setting is as small as possible.

Assumptions Underlying Experimental Design and Interpretation

Estimation of treatment effects boils down to comparing outcomes of units that get an intervention to those that do not. The key difficulty in measuring the effect of an intervention is to separate the treatment effect from other sources of difference across the treated and untreated subjects. In cases where the intervention is not randomly assigned, those who receive the treatment may be systematically different from those who do not. This problem, referred to as selection bias, is often common sense. In many cases the researcher is interested in situations in which people or organizations get to choose among different "treatments." These choices reflect a subject's resources and preferences, and whether a given treatment is desirable to the individual or organization is often determined, in part, by the same outcome variable that interests the researcher. For instance, people who like to exercise and place a priority on fitness may also join gyms. The observed average difference in fitness among those who belong to a gym and those who do not 7would not measure the effect of gym membership and would not serve as a reasonable forecast of a policy that provides free gym memberships.

Random assignment can be used to overcome the selection problem. We formalize the problem of estimating the treatment effect to demonstrate what is gained by random assignment and to describe the important additional assumptions that are necessary for an experiment to produce unbiased estimates of the treatment effect of interest to the researcher. We also discuss two related issues of generalization beyond the experiment: first, do the results of the experiment generalize to other groups beyond the experimental subjects, and can the experiment be generalized to interventions that are similar to the experiment but not identical to it? Second, do the experimental results apply to larger scale versions of the experimental intervention?

A Potential Outcomes Model of Causal Effects

Suppose the researcher wants to measure the effect of a program or policy. For concreteness, suppose the researcher is interested in measuring the effect of health insurance coverage on healthcare utilization. For each unit of observation, which we will call a "subject" even if it happens to be an organization, city or other entity, we measure whether the subject is "treated" or not. To be clear, although this monograph will focus on experimental work, the terminology of "treatments" and "outcomes" applies equally to experimental research and observational data. We denote the treatment status of subject i by D_i, and D_i=1 if i is treated and 0 if untreated, regardless of whether treatments are assigned randomly. In our example, let D_i=1 if unit i has good health insurance, and 0 if subject *i* has poor coverage or no insurance. The outcome of interest, in this example, is healthcare utilization, and we denote the outcome for subject *i* by Y_i. Datasets will often contain covariates, measuring other things about the subject beyond treatment status and the outcome. We ignore the other measured (pre-treatment) traits of the subjects for now; the discussion can be thought of as describing analysis within blocks formed by grouping subjects with similar background characteristics.

For each subject we define a pair of "potential outcomes," Yi(Di), which are the values of the outcome variable *Y* that would be observed in two circumstances: when the subject is treated, $Y_i(D_i=1)$, and when the subject is untreated, $Y_i(D_i=0)$. Although both of these potential outcomes are real quantities that could in principle be measured, we will actually observe only one of these potential outcomes, depending on whether the subject is treated or not. The treatment effect for an individual, in our example the difference in healthcare usage with and without good insurance, is defined as $Y_i(1)$-$Y_i(0)$. The average treatment effect for the collection of subjects, the sample average treatment effect (SATE), is $E(Y_i(1)-Y_i(0)$.

Random Assignment and Selection Bias

From the data available for our collection of subjects, we can calculate the difference in the average values of Y for the treated (those with good insurance) and the untreated (those without good insurance):

(1) $E(Y_i(1)|D_i=1) - E(Y_i(0)|D_i=0)$,

where the notation E(Ai|Di=B) means the average value of Ai among those subjects for which the condition Di=B holds.

In our example, this quantity is the difference in average healthcare use by the well insured versus use by those with poor or no insurance. Unfortunately, this is not equal to the average treatment effect from having "good versus bad" insurance. Rather, we learn the outcome for the treated subjects in their treated state (average healthcare usage for the insured) and the outcomes of the untreated subjects in their untreated state (average healthcare usage for the poorly insured). To see how this is different from the treatment effect (SATE), the expression (equation 1) can be rewritten as:

$E((Y_i(1)-Y_i(0))|D_i=1) + [E(Y_i(0)|D_i=1) - E(Y_i(0)|D_i=0)]$.

The difference in the average outcome of the treated and untreated can be decomposed into the sum of two quantities: the average treatment effect for a subset of the subjects (the treated), and a selection bias term. The selection bias term is the difference in what the outcome (Y) would have been for those who are treated had they been untreated and the value of Y observed among those who were not treated. This second term is an extra part of the difference in observed outcomes for the treated and untreated that is not the treatment effect. In the case of health insurance, among those with similar observables (for example age, class, race, etc.), standard economics arguments about adverse selection suggest that if future claims are, in part, anticipated by the individual but

unobservable to the insurer, then those with higher expected claims will be more likely to buy insurance. That is, even if they did not have insurance, the insured might use more medical care than the uninsured. In the case that a selection bias term is positive, the observed difference in healthcare usage among the insured versus the uninsured will over-estimate the treatment effect. More generally, however, it is difficult to know the magnitude of selection bias, and sometimes even the sign of the bias is unclear.

The selection problem is a solid obstacle to estimating causal effects from standard observational comparisons. Random assignment removes this obstacle. Intuitively, the difference in observed means across the treated and untreated is not equal to the average treatment effect because the quantities we measure are not the quantities we ideally want to measure. We observe group outcome averages among subjects who select into each group (treated or not) while wanting to obtain an average of a representative sample of the subjects in their treated state and untreated states. This is what random assignment provides. When random assignment is used to determine which group a subject is placed into, we know the process determining whether a unit is assigned to the treatment or not and we know (crucially) that the assignment is, by design, independent of the subject's potential outcomes. As a consequence, those randomly selected from the sample to form the treatment group are, except for the play of chance, representative selections from the group as a whole. This implies that, on average, they are expected to have the same treated potential outcomes as those randomly assigned to remain untreated (control group):

$E(Y(1)|D=1) = E(Y(1)|D=0) = E(Y(1))$.

By the same reasoning, those randomly assigned to the control group (to remain untreated) have, on average, the same outcomes in the untreated state as those assigned to the treatment group:

$E(Y(0)|D=0) = E(Y(0)|D=1) = E(Y(0))$.

There are two critical implications of random assignment. An implication of these equations is that when subjects are assigned using random assignment, the selection bias term vanishes and the difference of treatment and control group means measures the sample average treatment effect (SATE). This can be shown by substituting equations (3) and (4) into equation (1):

$[E(Y_i(1)\,|\,D_i=1) - E(Y_i(0)\,|\,D_i=1)] + [E(Y_i(0)\,|\,D_i=1) - E(Y_i(0)\,|\,D_i=0)] =$

$[E(Y_i(1)\,|\,D_i=1) - E(Y_i(0)\,|\,D_i=1)] =$

$E(Y(1)) - E(Y(0)) = E(Y(1)-Y(0))$.

Random assignment solves the selection problem, which is why it is such an attractive research method. However, random assignment must be supplemented by two additional assumptions if we are to draw meaningful causal inferences. We next discuss these two important assumptions.

Key Assumptions for Measuring the Effect of the Intended Treatment

Exclusion Restriction. First, the effect of a subject being assigned to the treatment group (or the control group) must be produced by the treatment itself rather than through some other channel that accompanies a subject's group assignment. To discuss this issue precisely, we need to introduce additional notation. Let $Z_i=1$ if a subject is assigned to the treatment group, and $Z_i=0$ if the subject is assigned to the control group. The exclusion restriction assumption is satisfied when:

$Y_i(D_i,Z_i)=Y_i(D_i)$ all D_i, Z_i, where $D_i=1$ if the subject is treated and 0 otherwise.

In the simplest experiments, all of those assigned to the treatment group (Z=1) also receive the treatment (D=1) and all those assigned to the control group (Z=0) do not receive the treatment

(D=0). This suggests that the exclusion restriction is an assumption and not something that can be empirically investigated.

There are two main ways that the exclusion restriction is violated. First, the treatment designed by the researcher may combine the channel of interest, D, with some other "active" ingredient. If so, the experiment provides an unbiased estimate of the treatment combination, but it does not estimate the separate effect of the portion of the treatment that embodies the principle motivating the experiment. The Hawthorne effect, whereby people act differently when being observed by others, may be thought of as an exclusion restriction violation, if the researcher is interested in the posited effect of the program rather than the incidental effect of being observed. Placebo effects can be considered exclusion restriction violations as well, if the control group gets no treatment rather than a placebo treatment. As an example of compound treatments, consider an experiment in which applicants are selected by lottery to receive health insurance. Let Z_i equal 1 if the subject wins the lottery, 0 otherwise. Suppose that the researcher wishes to interpret the treatment effect as the effect of providing insurance, but the actual treatment that is implemented combines health insurance with other forms of support and assistance, such as health counseling and information about other government health programs that might be of interest to the lottery winners. Let D_i=1 if a subject gets health insurance, 0 otherwise. Note that assignment to the treatment group involves much more than just the health insurance, which is just part of the collection of things that follow the lottery win. In this example, the difference of treatment and control group means is not an unbiased estimate of the treatment effect of the insurance alone. To link this to the formal statement of the exclusion restriction, Y(0,1) is the (theoretical) healthcare usage when the subject does not get health insurance but gets everything else a lottery winner gets, and Y(0,0) is usage when the subject does not get health insurance and is a lottery loser. Y(0,1) does not equal Y(0,0) if the health counseling and information about other programs has an effect on the

outcome, Y. If the researcher is interested in measuring the effect of the "full package" rather than a component of it (the insurance alone), then there is no violation of the exclusion restriction.

More subtly, any aspect of the experimental design that disturbs symmetric handling of the treatment and control group threatens to cause biases. In contrast to combination treatments, symmetry violations are typically inadvertent. Asymmetric measurement, for instance, may produce biased results. For example, suppose that all subjects in a health insurance experiment are called every month and asked about their healthcare usage. In addition, for those in the treatment group, the survey reports are augmented by the records generated whenever a subject uses his or her insurance card. This asymmetry will create bias, since measurement error will differ across treatment and control groups. The basic injunction both in design and measurement is to avoid *anything* that creates a difference in the treatment and control group outcomes through some channel other than the intended treatment.

The exclusion restriction may be especially important in cases where a substantial portion of those assigned to the treatment group do not actually get treated. When subjects assigned to a group do not receive what is intended, we say there is non-compliance (that is, $Z_i=1$ but $D_i=0$, or $Z_i=0$ but $D_i=1$). Non-compliance is common in field experiments. For many experimental designs, especially those involving government programs, subjects randomly assigned to the treatment group (that is, for whom $Z_i=1$) have won a lottery granting eligibility for a benefit, such as a private school voucher or an insurance program. Of those who win eligibility, only a fraction exercise their option to enroll in the program and therefore many in the treatment group remain untreated (they do not receive a private school education or the insurance program). In the simplest case of non-compliance, such as the example just provided, the non-compliance is "one-sided," which means that some of those assigned to the treatment group fail to be treated, while all of those assigned to the control group remain untreated. In such cases, the average treatment effect from the program

among those who take the treatment if assigned is calculated by dividing the difference between the average treatment group outcome (the entire treatment group, including those who were not treated) and the average control group outcome by the proportion of the treatment group actually treated. The observed difference between the treatment and control group averages is inflated by dividing the difference in group averages by the fraction actually treated. For example, if 1/3 of those who are a randomly granted eligibility for insurance actually receive the insurance, then the observed difference between the average outcome for the entire treatment and control is multiplied by 3 in order to estimate the effect of the treatment on those assigned for treatment.

To see how a violation of the exclusion restriction might affect treatment effect estimates when there is substantial non-compliance, suppose that winners of an insurance eligibility lottery also receive information encouraging subjects to purchase the insurance, such as information about the importance of regular medical exams. Then all of those in the treatment group (all the lottery winners) get this information whether or not they ultimately buy insurance. If this additional information accompanying treatment group assignment has the effect of boosting the average healthcare utilization by w, it will boost the estimated complier average causal effect by w/c, where c is the fraction of the treatment group that enrolls in the program. If c is small, such as 1/5 or 1/10, then the bias will be $5w$ or $10w$, showing how even a small violation of the exclusion restriction can lead to substantial bias in the estimated treatment effect.

Finally, consider the somewhat more complicated case of two-sided non-compliance. Two-sided non-compliance occurs if some subjects assigned to the treatment group are not treated and some subjects assigned to the control group nevertheless receive the treatment. Since access to the treatment is under the control of the researcher in many experimental designs, two-sided non-compliance is impossible. However, two-sided non-compliance arises naturally in some types of experiments. In encouragement designs, for instance, the treatment group is given an added

inducement to obtain a treatment. An example of this would be a public health program that encourages subjects to get their children vaccinated. Many in the control group will vaccinate their children, and some in the treatment group will not.

Another design in which two-sided non-compliance arises is downstream experiments, which are analyses in which the outcome measure in a first experiment is recast as the treatment in a subsequent investigation. These analyses are quite common in the study of health outcomes. An example is the frequently studied natural experiment conducted during the Vietnam War era draft lottery. To simplify, young men randomly received either a high or low draft number. For each of these randomly formed groups, a proportion, H or L, H > L, subsequently served in the war. The treatment effect of the lottery on whether the subject served in the military (Y=1 if served in military, 0 otherwise) is estimated by the observed proportion H – L, and the lottery is an experiment absent of non-compliance (the treatment is the lottery number and all subjects received the lottery number they were assigned).

Now suppose the researcher is interested in measuring the effect of military service on health outcomes. This may be thought of as a "downstream" experiment because the initial randomization has two sequential effects: an effect on military service and a subsequent effect on another outcome, health status. The setup shifts to accommodate this new use of the random assignment: Z_i is the random assignment to experimental group (the lottery group), D_i (not Y) denotes military service, and Y_i measures health outcomes. Note that we also have a situation with two-sided non-compliance: some of those assigned to the control group (Z_i=0, the low lottery number group) serve in the military (D_i=1), and some of those assigned to treatment (the high number group) do not serve.

Measuring treatment effects when there is two-sided non-compliance requires modification of the formula used for the case of one-sided non-compliance. In order to estimate the average

treatment effect among those who take the treatment if and only if they are assigned to the treatment group, we now take the average difference in Y for the treatment (Z_i=1) and control group (Z_i=0) and divide by the difference in the treated proportion of each group. In this example, the average difference in health outcomes among those assigned the high and low lottery numbers would be divided by H-L. We can now reflect on whether the downstream experiment satisfies the exclusion restriction. To interpret this ratio as an estimate of the net effect of military service on health, it must be the case that lottery group assignment does not have an effect on health outcomes through some channel other than military service. For instance, if those with a high draft number took measures to avoid induction, including leaving for Canada, changing marital or education statuses, or shooting off their toes, and these affected subsequent health outcomes, then the health effects estimates would be biased.

Non-interference. A second key assumption is non-interference across units. This assumption requires that each subject's potential outcomes, $Y_i(1)$ and $Y_i(0)$, are not affected by which other subjects are selected at random to be in the treatment versus control group. Notice that this assumption is already incorporated into the notation we have used for the potential outcomes, since Y is written as a function of Z_i (or D_i) and not as a function of the treatment or control group assignments of any other subjects. A leading example of how this assumption might fail is when there are "spillover" effects from treatments. For instance, suppose the treatment is a flier informing subjects of a concert and the outcome is attendance at the concert, or Y_i. If the flier induces conversations about the concert and some control group subjects thereby learn of the concert from their treated friends, then treating subject i (assigned to the treatment group) has affected subject j (control group) in the untreated state. If subjects in the control group show up to the concert due to the spillover treatment, then this will raise the average outcome in the control

group and lower the difference between the treatment and control group average. Interference has caused the true effect of the treatment to be underestimated.

When non-interference is plausible, the assumption permits us to ignore any effects of the treatment given to any other subject on subject i. If this assumption does not hold, the distinction between the treated and untreated begins to breaks down, since a subject may be "untreated" but nevertheless his or her outcomes may be affected by the treatments being dispensed to others. If so, the untreated subjects do not serve as a clean counter-factual for the treated in their untreated state. Back to our example, if providing more financial support for health insurance for subjects in the treatment group causes government to cut back its health expenditures on behalf of the control group, the non-interference assumption will be violated. In that case, a comparison of average outcomes in the two groups no longer gives an unbiased indication of the effect of receiving health insurance because the control group is not untreated; instead, the control group receives an adverse treatment.

The non-interference assumption is quite natural in medical interventions, since there are rarely channels for the treatment applied to one patient to have a material effect on the experience of another patient. In the context of social experiments, the plausibility of non-interference must be evaluated on a case-by-case basis. Communication, contagion, social comparisons, or displacement may cause potential outcomes for a given subject to depend on the treatments administered to other subjects.

When the non-interference assumption is false, the potential outcomes model, as written, does not accurately reflect the situation the researcher is trying to represent. There are various approaches to address interference between units. First, studies may be designed to measure the extent and pattern of interference. For example, randomly varying the density of treatment across neighborhoods and comparing those assigned to the control group across neighborhoods will

measure geographic spillover within neighborhoods (Angelucci & De Giorgi, 2009; Duflo & Saez, 2003; Hirano & Hahn, 2010; Lalive & Cattaneo, 2009). Second, although the potential for contamination may prevent the measurement of individual-level treatment effects, changing the unit of observation (typically choosing a unit of observation more coarse than the individual) may reduce or eliminate the danger of spillover. If so, the non-interference assumption is satisfied for the experiment when the researcher randomizes the coarser units. To illustrate this concept, it is not difficult to imagine that the complex interactions between students, and between students and teachers, may make it impossible to measure the effect of an alternative curriculum provided to a randomly selected subset of a class. However, randomly assigning whole classes (or perhaps whole schools) to different curricula may arguably satisfy the non-interference requirement and therefore provide a credible estimate of intervention effects. Murray (1998) provides extensive guidance on the statistical analyses of such group-randomized trials.

Interpretation and Extrapolation of Experimental Results

Finally, interpreting and extrapolating findings from a randomized trial requires care. Regarding interpretation, the measured experimental outcomes are often proxy variables for the true outcome of interest. This is very common in medical studies, where the severity of illness and mortality is the true outcome variable of interest, but experimental treatments are evaluated based on more proximal indicators such as blood test results or changes in tumor size, which are assumed to be related to more distal outcomes of interest such as reduced early mortality. The validity of any conclusions about the effect of the treatment on the true outcome of interest then rests upon the confidence with which the research can move from the proxy to the ultimate outcome of interest. In law or policy interventions, a key challenge is that subject behavior may change along many dimensions in

response to an intervention. If what is measured does not provide the whole story, the results of the experiment may be misleading. For example, a recent study of the effect of advertising looked at the change in sales associated with increasing the number of clothing catalogues mailed to households by a major retailer (Simester et al., 2009). Compared to a control group that received the normal flow of mailings, the households getting the additional catalogues ordered a substantially greater amount of clothes during the period of the experiment. Based on this sales boost, additional mailings appeared to be a profitable idea. However, the researchers also measured household internet purchases and purchasing behavior during the year after the experimental mailings had concluded. It turned out that the treatment group decreased internet purchasing and also purchased less than the control group during the year after the experimental mailings ended. Once channel and intertemporal substitution were measured, the overall effect on sales was not large enough to cover the cost of the additional mailings.

Substitution can also occur in response to health interventions. Consider an experiment to measure the effect of posting calorie counts on lunch menus in a random selection of fast food restaurants. It is useful to know if the intervention results in more salads being ordered at treatment group restaurants during the typical lunch hour. But if the quality of the public's overall diet is the variable of interest, it would be important to know how posting calorie counts at lunch affects what people eat at dinner and whether a light lunch leads to larger afternoon snacks. Further, if overall health rather than diet is the outcome of interest, researchers should attempt to measure whether ordering in a "virtuous" way at lunch leads the subject to skip exercise in the evening. Of course, oftentimes behaviors are complements, and eating a salad at lunch might encourage the subject to build on his earlier good behavior and exercise. The general point is that if the experiment has only limited outcome measures, and other unmeasured behaviors of significant relevance to the outcome

of interest are substitutes or complements, caution is in order when interpreting the experimental results.

A common use of experiments is to predict the effects of similar interventions applied to new subjects or in new contexts, or the effects of the same intervention scaled up from a small program to general policy. It is critical to remember that when all goes well what is obtained from an experiment is the average treatment effects for particular subjects in a specific context in response to the treatment used. Extrapolating moves us from the relatively firm ground of unbiased average treatment effect estimates secured through random assignment to theoretical conjectures about what the effects would be with modified treatments applied to a different set of subjects in other times and places. Unless the experiment includes variation in treatment or context, any extrapolation along these dimensions involves substantial guesswork, and the quality and reliability of these translations will vary from case to case. Subjects' attributes may also limit generalizability of findings when subjects are not selected randomly from some larger population. For instance, suppose the subjects in an experiment in which insurance eligibility is determined by a lottery must have very low incomes and few assets to participate in the lottery. It is possible that the results apply well to any policy aimed at the poor, but do not apply to middle-class uninsured. Because random sampling is relatively unusual in field experiments, generalizing from sample average treatment effect to population average treatment effect requires the researcher to impose additional assumptions (see Aronow & Sovey, 2011).

In order to shorten the distance between experimental intervention and real-world program, researchers sometimes design "place-based" experiments in which communities, school districts, or regions are assigned as units to treatment or control. Although this design sacrifices some of the statistical power that comes with the random assignment of large number of individuals, it has the advantage of allowing the researcher to evaluate a realistic package of treatments using natural

administrative units. School districts, for example, might ordinarily administer a healthy eating initiative, and assigning entire school districts to this intervention allows researchers to study the net effect of an entire school district – students, teachers, administrators, and parents – working together in pursuit of this goal. See Boruch (2005) and Bloom (2009) for reviews of place-based studies across a wide-array of disciplines.

Even when one's evidence base consists of place-based studies, there are also reasons for caution when using the results of a small-scale intervention to predict the effects of a large-scale version of the program. While third parties will ignore a small pilot study, a larger scale intervention (or interventions which look to be long term and not just "pilot" programs) may trigger responses from individuals, groups, or organizations not directly treated by the intervention but who are affected by the program. For example, a small-scale private school voucher experiment may be ignored by the local public schools, but a large-scale voucher program might trigger a competitive response, such as an effort to improve the public schools. This potential effect of the large-scale intervention will go unmeasured by the initial randomized experiment. Similarly, if a substantial percentage of the uninsured in a neighborhood are given insurance, this may lead to crowding of local medical resources, unless additional clinics or practices open in response to the program. In addition, holding the actions of third parties fixed, the effect of the treatment may vary with the scale of the program. In a neighborhood in which everyone has health insurance, residents may discuss medical appointments and local doctors as topics of common interest and so information might spread freely. This may not be the case when only a few insured "lottery winners" live in a neighborhood with a low rate of insurance participation. By expressing this note of caution, we do not mean to set an impossibly high standard of evidence; rather, our point is that answering big policy questions requires thoughtful designs and extensive replication across an array of settings.

Applications of RCTs to Law and Health

RCTs have become pervasive in program evaluation and social science investigations of policy-relevant theories. In this section, and in Table 1, we provide readers with a sense of how RCTs have been used in five areas that are directly or indirectly related to health: institutions, disease prevention, education, migration, and healthcare.

Institutions

A common critique of RCTs is that they cannot be used in many issue areas due to difficulty I logistics and practice. For example, in research exploring effects of institutions on health and development, it could be argued that independent variables of interest simply cannot be randomized. At first glance, an analysis of the effects of variables like democracy, type of institution, property rights, civil liberties, and corruption on health outcomes certainly seems impervious to experimental manipulation. As a result, researchers have produced a plethora of observational research designs in this area, but these have been plagued with problems such as reverse causality, omitted variables bias, and measurement error. For example, democracy may cause development, or perhaps development causes democracy. These types of problems resulted in a failure to produce conclusive evidence (Boix & Stokes, 2003; Przeworski & Limongi, 1997).

Recently, scholars have begun to design clever RCTs in this area, which are able to overcome such problems through the use of random assignment. In a prominent study of the effect of democratic institutions on the choice of public health policy, Olken (2010) randomized forty-eight villages in Indonesia to use either an elite meeting or a village plebiscite to decide on two project proposals related to infrastructure and public health (roads/bridges, water/sanitation, health/education, or irrigation). One project was proposed by the village and the other was

proposed by only the women in the village. Outcome measures were the closeness of proposed projects to elite preferences, proximity of proposed projects to wealthier areas of the village, and villager satisfaction with the proposals. Olken found that while the treatment had little effect on projects proposed, it had a large effect on villager satisfaction. He concludes that direct democracy can increase satisfaction and legitimacy. Olken (2010) demonstrates that variables previously thought out of reach, such as democracy, can be randomly assigned.

Other interesting examples of the randomization of institutional variables abound. Communities in post-conflict Liberia were randomly assigned to receive a community-driven reconstruction program to determine the effect of the program on money raised for a collective project (Fearon et al., 2009). The researchers conclude that these types of institution building projects can increase social cohesion. For comprehensive summaries of this literature see Humphreys and Weinstein (2009), De La O and Wantchekon (2010) and Moehler (2010).

Disease Prevention

RCTs have been used extensively in the disease prevention literature. RCTs have been important, for example, for the debate about whether policy-makers should charge fees for preventive services. On the one hand, those who advocate charging fees argue that fees increase the value of the service in the minds of consumers, people may be more likely to use the product because they have already sunk a cost in it and fees may encourage providers to provide the service. Thus, many organizations, such as the non-profit social marketing organization Population Services International, charge fees for items such as mosquito nets and water disinfectant (Population Services International, 2003). On the other hand, other organizations like the World Bank have moral qualms about charging for health products and therefore tend to give products away (World Health Organization, August 16, 2007).

To help resolve these conflicting views, RCTs have been run in many countries to assess effects of charging for a variety of health products. These studies conclude that usage of health products is highly responsive to price. For example, Kremer and Miguel (2007) investigated the effect of charging for a deworming program in schools. In a prior study, Kremer and Vermeersch (2004) showed that deworming programs boost school attendance and prevent worm infections. Therefore, understanding the factors that cause students to participate in the program is important. They randomly assign schools to share in the cost of the program at an average rate of thirty cents per student, and find that seventy-five percent of students assigned to free treatment schools participate while only nineteen percent of students assigned to cost sharing schools participate. Other RCTs showing that lower prices lead to greater usage include studies of the effect of cost changes for mosquito nets (Cohen & Dupas, 2010) water disinfectant (Ashraf et al., 2007), HIV results and counseling (Thornton, 2005) and vaccines (Banerjee et al., 2010). (See Holla et al., 2009 for a review.)

RCTs have also been used to evaluate alternative intervention strategies. Access to clean water is a central public health challenge in the developing world and policy makers must sometimes decide whether to spend limited resources on efforts to improve the quality of water or increase the availability of water. Observational studies have a difficult time distinguishing between the marginal benefit from increasing quantity versus quality in water provision (Gamper-Rabindran et al., 2010; Watson, 2006). Based on observational research, many researchers believed that improving water quantity was crucial because it would encourage subjects to wash hands and bathe more frequently (Curtis et al., 2000; Esrey, 1996).

Recent RCTs however, have called this finding into question. For example, Devoto and colleagues (2011) find that increasing quantity without altering quality has no effect on health outcomes. Among villagers in Northern Morocco without access to tap water in their homes, the authors randomly assigned an offer to buy a household connection to tap water on credit. The tap water was of the same quality as the water which was publically available to all villagers, so the intervention altered the quantity of water without changing the water quality. Although sixty-nine percent of treated households were willing to purchase the household tap water access, the authors found no change in the health of the subjects from the intervention. Even as questions of external validity remain, further RCTs are a promising avenue for research in this area.

Though RCTs have cast doubt on the hypothesis that water quantity is crucial for improved health, they have confirmed that improved quality, through filtration or chlorine treatment, can improve health quite dramatically (see Ahuja et al., 2010; Harrell & Smith, 1996; Waddington & Snilstveit, 2009 for a review of this literature). However, RCTs have also shown that subjects may not be willing to pay much for increased water quality. Kremer and colleagues (2009a) and Holla and colleagues (2009) randomly assign discounts for chlorine treatment to villagers in Kenya and show that purchase of the treatment is highly sensitive to price. While over half of households use it when it is delivered to their home for free, less than ten percent use it when it is only thirty cents a month. Demand changes little when households receive half price coupons, have young children, or know a peer who uses the treatment (although hiring local promoters of the treatment does boost usage). Additionally, Kremer and colleagues (2009a) show that there is little evidence of spillover. That is, knowing someone who uses the system does not appear to increase usage. Other RCTs have confirmed Kremer's findings (Ashraf et al., 2011), adding that information effects are small compared to price effects (Jalan & Somanathan, 2008).

The conclusion from this strand of literature seems to be that although health is most improved by increasing water quality, households are most willing to pay for increased water quantity. Taken together, this body of research shows how initial RCTs and subsequent follow up experiments can produce a more nuanced view of the relative strengths of alternative public health policies.

Education

Recent research has strengthened the positive link between educational attainment and individual health (Cutler & Lleras-Muney, 2006). Randomized trials have been used to evaluate alternative strategies for improving schools, including requiring remote schools in Africa to provide photographic evidence of school and teacher attendance (Duflo et al., 2008), and paying teachers and schools various bonuses if their students do well on standardized tests (Fryer, 2011). A hotly debated topic in this area is the effect of school vouchers on improving school quality. Vouchers fund students to attend schools other than those they would be assigned based on geographic proximity. Voucher programs have become increasingly popular and vary in their design and administration; programs differ in the types of students they favor, the way in which they are funded, the amount of aid given and the period during which aid is given. Examples of publically funded programs include: the EdChoice Scholarship Program in Ohio, the A+ Opportunity Scholarship Program in Florida, the Cleveland Scholarship and Tutoring Program in Ohio and the Milwaukee Parental Choice Program in Wisconsin. Examples of privately funded programs include the Washington Scholarship Fund in Washington DC and the School Choice Scholarships Foundation in New York.

While observational studies suggest that voucher programs are highly effective, confounding factors may bias the results. Although students self-select into private schools, many studies rely on comparisons between private and public school students. If students who self-select into private schools differ systematically from those who do not, the results will be biased. Other observational studies which compare voucher recipients to non-recipients can be similarly confounded because students who apply for vouchers may be more motivated than those who do not apply.

To overcome potential biases, analysts have made use of privately funded voucher programs whereby applicants are randomly selected to receive vouchers. Because vouchers are randomly assigned to applicants, the pool of applicants who apply for vouchers and receive them should not differ systematically from the pool of applicants who apply for vouchers and do not receive them. In contrast to observational studies, experimental studies typically find limited effects of vouchers on test scores, though parents of students who received vouchers reported increased satisfaction with their children's educations and felt their schools were safer (Howell & Peterson, 2002; Wolf et al., 2010).

Another closely-related debate in the education literature is the effect of educational subsidies. Mexico's Programa de Educacion, Salud y Alimentacion (PROGRESA) was a pioneer in this area. The program gave poor mothers cash grants if their children attended school eighty-five percent of the time, and increased the amount of the grant with the child's grade level in school. Cash grants were also given for participation in a variety of health and nutrition programs. For the first two years, the program was administered in randomly selected poor areas of the country, after which it was expanded to all regions. The use of random assignment allows researchers to evaluate the program, which was found to increase enrollment by 3.4 to 3.6 percentage points in grades one through eight and increased total schooling by 0.66 years (Schultz, 2004). Based on the success of the program, similar programs have been implemented in almost 30 other countries, many of which

use random assignment. In addition to cash transfers, merit-based subsidies have been found to increase enrollment and attendance when given for school meals (Kremer & Vermeersch, 2004), school uniforms (Kremer et al., 2003) or school fees (Kremer et al., 2009b).

Migration

Researchers in the social sciences are interested in the impact of migration on a variety of outcomes, including health, nutrition, and income, and in designing migration policies that promote these outcomes. However, studying effects of migration is inherently difficult due to the fact that migrants typically choose whether to migrate. If those who choose to (or are able to) migrate are different from those who do not, estimates from observational research may be biased. RCTs remedy this problem by using random assignment. One source of randomization that has been effectively exploited is Visa lotteries, which randomly select applicants for receipt of Visas. Each year, the Pacific Access Category program in Tonga randomly chooses 250 applicants to relocate to New Zealand. By comparing those randomly chosen to migrate with those randomly chosen not to migrate, researchers have found that the average income of those who migrate increases by 263 percent within one year of moving (McKenzie et al., 2010) and that mental health of migrants improves (Stillman et al., 2009). Interestingly, McKenzie and colleagues (2010) used a sample of the total population and compared estimates from their RCT to those they would have obtained from an observational study and found that the observational study would have exaggerated the income effect by 27-35 percent. Other policy experiments have exploited randomized designs in clever ways. Gibson and colleagues (2010) used the rule that a person selected to migrate often may only bring a spouse to assess the effect of migration on remaining household members. They found that remaining members had lower incomes and consumption. Clemens (2010) looked at effects of migration on a particular firm by using the lottery for an H1-B visa designed to admit high-skilled

workers into the United States. However, as McKenzie and Yang (2010) point out, if wages for non-migrants within the firm increase as a result of migrants leaving, then non-interference would be violated in the Clemens study.

Rather than rely on governments to randomize policy implementation to individuals, organizations or communities, some researchers study migration by conducting their own RCTs. For example, Ashraf and colleagues (2011) offered Salvadoran migrants who migrated to Washington D.C. the opportunity to exercise control of their remittances by channeling them into savings accounts in El Salvador. The researchers randomly varied the amount of control over the accounts, so that migrants were either offered savings accounts owned by their spouses, owned jointly, or owned by the migrant. They found that the degree of control had a large effect on remittances, and are currently investigating effects on other outcomes such as health and nutrition. Researchers are pursuing other RCTs in this area, such as Chin, Karkoviata and Wilcox (2010) who randomly offer an ID card that can be used at United States' banks to Mexican migrants. The ID card leads migrants to open more savings accounts in the United States and to send fewer remittances home. Additionally, some scholars are now running RCTs to study barriers to migration by randomizing the provision of information about jobs in the host country, cash or credit assistance, or assistance in filling out applications (Beam et al., 2010; Bryan et al., 2010).

RCTs exploring the effects of migration have looked at both domestic and international migration. On the domestic side, a major area of research has explored effects of neighborhood income level on physical and mental health. For example, researchers assessed the effect of the Moving to Opportunity program (MTO), which was implemented in five major cities in the United States. Families living in public housing were randomly assigned to one of three groups: the control group remained eligible to live in public housing, the first treatment group received vouchers to move to their neighborhoods of choice and the second treatment group received vouchers to move

to low income neighborhoods. Large effects on mental and physical health are reported (Kling & Liebman, 2005 ; Liebman et al., 2004).

Healthcare

Studying the effects of healthcare laws and regulations can be difficult since subjects participate in the choices; those choosing the lowest amount of insurance may be the healthiest subjects. RCTs have been able to overcome this obstacle. For example, to investigate whether receiving greater health insurance coverage increases obesity, researchers analyzed data from the RAND Health Insurance Experiment, which was conducted in the early 1970s in six regions in the United States. The experiment randomized subjects to receive varying amounts of health insurance coverage and found that more insurance coverage had no effect on body weight (although moving from uninsured to insured may have increased body weight) (Bhattacharya et al., 2011). Similarly, data from the U.S. Working Toward Wellness program revealed that randomized encouragement to seek mental health services had mixed results on depression severity (Kim et al., 2009). Or, as another example, low-income people who were randomly assigned opportunities to apply for Medicaid were found to have higher self-reported mental and physical health, more frequent use of health facilities and lower health-related expenditures (Finkelstein et al., 2011).

Healthcare RCTs have not been confined to the United States. In 2003, the Mexican government passed a modification to Mexican health policy called "Seguro Popular" which provides free and subsidized healthcare to lower income Mexicans. To analyze the effectiveness of the program, Mexico was divided into 12,284 regions called "health clusters". Health clusters were randomly assigned to receive encouragement to participate in the program in 13 of the 30 Mexican states where the government agreed to participate in the experiment. Outcome measures included survey responses indicating how citizens felt about the quality of their healthcare as well as actual

measures of health such as blood pressure (King et al., 2007). A randomized trial in Uganda found community-based monitoring of healthcare providers improves quality of healthcare provision (Björkman & Svensson, 2009).

Another interesting strand of research that falls into this category include the RTCs that investigate the psychology behind how people choose to care for their own health. For example, Wansink and colleagues (2006) have conducted a variety of experiments whose overall implication is that the amount of food people eat is greatly influenced by factors other than hunger. His innovative experiments include random assigning: the size of popcorn buckets for movie-goers, whether diners receive bottomless soup bowls, whether diners are allowed to refill their plates or whether diners are told their wine is from North Dakota or California. If people made decisions about the amount of food to consume based solely on their hunger, these seemingly unrelated factors should have had no influence on their food intake. However, those who received larger popcorn buckets, bottomless soup, were allowed to refill their plates and were told they received California wine all ate significantly more food (for summaries of these experiments see Wansink, 2006).

Or consider the series of experiments which are designed to assess the effects of public service announcements on behavior (Banks et al., 1995). The researchers investigated a variety of message frames encouraging mammography utilization (Banks et al., 1995), tobacco smoking cessation (Schneider et al., 2001), sunscreen usage (Detweiler et al., 1999) and improved diet (Williams-Piehota et al., 2004). A prominent finding to come from this line of research is that gain-framed messages were more effective in promoting healthy behaviors like sunscreen usage, while loss-framed messages were more effective in promoting detection of problems, like mammography utilization. The policy-relevance of this broad conclusion is bolstered by the fact that this hypothesis has held up in a wide array of settings.

Conclusion

The randomized trial is widely accepted as the most reliable method for measuring causal effects. A striking development of the past decade is the rapid spread of randomized experiments from medical and pharmaceutical trials to the social sciences, public health and beyond. It is now common to exploit naturally occurring random assignments or design experimental interventions, and this shift in research strategy has enormous implications for how we assess causal claims about the effects of law and policy. As demonstrated by the brief review of some recent literatures, an experimentally based literature on the effect of alternative institutions and policies is not just theoretically possible, but is experiencing active development.

Our brief review of a portion of the emerging experimental literature shows how pioneering applications of experimental methods to the study of public policy have already produced important achievements. Nevertheless, some caution is in order. Although random assignment overcomes one of the most significant barriers to evaluation of an intervention, namely that the intervention may be correlated with observable and unobservable differences related to the outcome, the results of experiments must still be interpreted with care. For the reasons outlined above, it is important to keep in mind how difficulties such as interference between experimental units, or violations of the exclusion restriction, may affect estimates from RCTs. To be sure, this is not an argument against random assignment or any reason to dampen enthusiasm for the method; standard observational research typically has all of the difficulties experiments face, and the selection problem as well.

The growth and development of experimental research in the social sciences may be viewed as a process by which threats to inference attract attention and inspire increasingly sophisticated and robust research designs. Concerned that experiments that administer deworming treatments individually to Kenyan schoolchildren might produce misleading results if applied to entire schools, researchers have studied the effects of village-level interventions (Miguel & Kremer, 2004).

Concerned that the health effects of cash transfer programs might be underestimated if one focuses solely on the low-income beneficiaries, researchers have investigated the downstream effects on those living in proximity to low-income beneficiaries (Schultz, 2004). Much the same goes for the investigation of policy-relevant mechanisms. A policy of free distribution of disinfectants and anti-malaria products may provoke critics who argue that the psychology of "sunk cost" means that people will not value and use products that they do not pay for. In response, a series of field experiments have been launched to assess both the sunk cost hypothesis in general and the usage and subsequent purchase of health products in particular (Ashraf et al., 2011; Cohen & Dupas, 2010). The broader point is that researchers are becoming increasingly adept at formulating experiments that address threats to core assumptions, replicating experiments in ways that address concerns about generalizability and crafting experiments in ways that illuminate the role of policy-relevant causal mechanisms (see Ludwig et al., 2011).

Table 1: Examples of RCTs in Health Law and Policy

Issue Area	Citation	Manipulation	Outcome Variable
Institutions	Olken (2010)	Direct Democracy	Proposal type and satisfaction
Institutions	Bertrand et. al. (2007)	Individual vs Social incentives	Bureaucratic Response
Institutions	Humphreys and Weinstein (2009)	Reconstruction programs	Money raised
Disease Prevention	Kremer and Miguel (2007); Cohen and Dupas (2010); Ashraf et. al. (2010, 2011); Thornton (2005); Banerjee et. al. (2010); Kremer et. al. (2009)	Price of deworming program; water disinfectant; mosquito nets, chlorine treatment; vaccine, HIV counseling	Purchase, usage, health
Disease Prevention	Deveto et. al. (2011)	Access to tap water	Health outcomes
Disease Prevention	Jalan and Somanathan (2008)	Information about water quality	Purchase
Education	Howell and Peterson (2002); Mayer et. al. (2002); Krueger and Zhu (2004)	Voucher allocation	Test scores, satisfaction, feeling safe
Education	Schultz (2004)	Cash grants for education	Total schooling, enrollment
Education	Kremer and Vermeersch (2004); Kremer et. al. (2003); Kremer et. al.(2008)	Subsidies for school meals; school uniforms; reduced school fees based on merit	Total schooling, enrollment, attendance
Migration	McKenzie et. al. (2010); Stillman et. al. (2009); Clemens (2010)	Visa lotteries	Income, mental health
Migration	Gibson et. al. (2010)	Migration policy rule	Income, consumption
Migration	Ashraf et. al. (2010)	Control over savings accounts	Remittances, health, nutrition
Migration	Chin et. al. (2010)	ID card for bank use	Savings accounts, remittances
Migration	Beam et. al. (2010); Bryan et. al. (2010)	Provision of information; cash or credit; assistance with migration process	Migration decisions
Migration	Liebman et. al. (2004); Kling and Liebman (2005)	Housing vouchers	Mental and physical health
Healthcare	Bhattacharya et. al. (2011)	Health Insurance Coverage	Obesity
Healthcare	Kim et. al. (2009)	Encouragement for mental health service	Depression severity
Healthcare	King et. al. (2007)	Healthcare subsidies	Health, satisfaction
Healthcare	Bjorkman and Svensson (2009)	Monitoring of health providers	Healthcare provision quality
Healthcare	Wansink (2006), Detweiler et. al. (1999)	Framing of health information, eating environment	Usage of health product, food intake

List of Tables

References

Ahuja, A., Kremer, M. & Peterson-Zwane, A. (2010). Providing safe water: Evidence from randomized evaluations. *Annual Review of Resource Economics, 2*(1), 237-256.

Angelucci, M. & De Giorgi, G. (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption. *American Economic Review, 99*(1), 486-508.

Aronow, P. & Sovey, A. (2011). From late to ate: Dealing with treatment effect heterogeneity in instrumental variables estimation. Unpublished Working Paper. Yale University.

Ashraf, N., Aycinena, D., Martinez, C. A. & Yang, D. (2011). Remittances and the problem of control: A field experiment among migrants from El Salvador. Unpublished Working Paper. University of Chile.

Ashraf, N., Berry, J. & Shapiro, J. (2007). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review, 100*(5), 2383-2413.

Banerjee, A. V., Duflo, E., Glennerster, R. & Kothari, D. (2010). Improving immunisation coverage in rural India: Clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *British Medical Journal, 340*(c2220).

Banks, S. M., Salovey, P., Greener, S. et al. (1995). The effects of message framing on mammography utilization. *Health Psychology, 14*(2), 178-184.

Beam, E., D., M. & Yang, D. (2010). Financial and informational barriers to migration: A field experiment in the Philippines. Unpublished Ongoing Study. University of Michigan and World Bank.

Bhattacharya, J., Bundorf, K., Pace, N. & Sood, N. (2011). Does health insurance make you fat? In M. Grossman and N.H. Mocan (Ed.), *Economic aspects of obesity* (pp. 35-64). Cambridge, MA: National Bureau of Economic Research.

Björkman, M. & Svensson, J. (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *The Quarterly Journal of Economics, 124*(2), 735-769.

Bloom, H. S. (2009). The core analytics of randomized experiments for social research. In P. Alasuutari, L. Bickman & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 115-133). Thousand Oaks, CA: SAGE Publications.

Boix, C. & Stokes, S. (2003). Endogenous democratization. *World Politics, 55*(4), 517-549.

Boruch, R. (2005). Preface: Better evaluation for evidence-based policy: Place randomized trials in education, criminology, welfare, and health. *Annals of the American Academy of Political and Social Science, 599*, 6-18.

Bryan, G., Chowdhury, S. & Mobarak, A. M. (2010). The effect of seasonal migration on households during food shortages in Bangladesh. Unpublished Ongoing Study. Yale University.

Chin, A., Karkoviata, L. & Wilcox, N. (2010). Impact of bank accounts on migrant savings and remittances: Evidence from a field experiment. Unpublished Working Paper. University of Houston.

Clemens, M. (2010). How visas affect skilled labor: A randomized natural experiment. Unpublished Working Paper. Tufts University, Center for Global Development.

Cohen, J. & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics, 125*(1), 1-45.

Curtis, V., Cairncross, S. & Yonli, R. (2000). Review: Domestic hygiene and diarrhea – pinpointing the problem. *Tropical Medicine & International Health, 5*(1), 22-32.

Cutler, D. M. & Lleras-Muney, A. (2006). *Education and health: Evaluating theories and evidence* (NBER Working Paper 12352). Cambridge, MA.

De La O, A. & Wantchekon, L. (2010). Experimental research on democracy and development. In J. N. Druckman, D. P. Green, J. H. Kuklinski & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 384-399).

Detweiler, J. B., Bedell, B. T., Salovey, P., Pronin, E. & Rothman, A. J. (1999). Message framing and sunscreen use: Gain-framed messages motivate beach-goers. *Health Psychology, 18*(2), 189-196.

Devoto, F., Duflo, E., Dupas, P., Parienté, W. & Pons, V. (2011). *Happiness on tap: The demand for and impact of piped water in urban Morocco.* (NBER Working Paper 16933). Cambridge, MA: National Bureau of Economic Research.

Duflo, E., Hanna, R. & Ryan, S. (2008). *Monitoring works: Getting teachers to come to school.* Cambridge, MA: National Bureau of Economic Research.

Duflo, E. & Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly Journal of Economics, 118*(3), 815-842.

Esrey, S. A. (1996). Water, waste, and well-being: A multicountry study. *American Journal of Epidemiology, 143*(6), 608-623.

Fearon, J. D., Weinstein, J. M. & Humphreys, M. (2009). Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia. *American Economic Review, 99*(2), 287-291.

Finkelstein, A., Taubman, S., Wright, B. et al. (2011). *The Oregon health insurance experiment: Evidence from the first year* (NBER Working Paper 17190). Stanford, CA: National Bureau of Economic Research.

Forster, J. L., Murray, D. M., Wolfson, M., Blaine, T. M., Wagenaar, A. C. & Hennrikus, D. J. (1998). The effects of community policies to reduce youth access to tobacco. *American Journal of Public Health, 88*(8), 1193-1198.

Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools.* (NBER Working Paper 16850). Cambridge, MA: National Bureau of Economic Research.

Gamper-Rabindran, S., Khan, S. & Timmins, C. (2010). The impact of piped water provision on infant mortality in Brazil: A quantile panel data approach. *Journal of Development Economics, 92*(2), 188-200.

Gibson, J., McKenzie, D. & Stillman, S. (2010). *Accounting for selectivity and duration-dependent heterogeneity when estimating the impact of emigration on incomes and poverty in sending areas* (Policy Research Working Paper 5268). Washington, DC: World Bank.

Harrell, A. & Smith, B. E. (1996). Effects of restraining orders on domestic violence victims. In E. S. Buzawa & C. G. Buzawa (Eds.), *Do arrests and restraining orders work?* (pp. 214-242). Thousand Oaks: SAGE.

Hirano, K. & Hahn, J. (2010). Design of randomized experiments to measure social interaction effects. *Economics Letters., 48*(1), 51-53.

Holla, A., Kremer, M. & Center for Global Development. (2009). *Pricing and access: Lessons from randomized evaluations in education and health* (Working Paper 158). Washington, DC: Center for Global Development.

Howell, W. G. & Peterson, P. E. (2002). *The education gap : Vouchers and urban schools*. Washington, DC: Brookings Institution Press.

Jalan, J. & Somanathan, E. (2008). The importance of being informed: Experimental evidence on demand for environmental quality. *Journal of Development Economics, 87*(1), 14-28.

Kim, S., LeBlanc, A. & Michalopoulos, C. (2009). *Working toward wellness: Early results from a telephone care management program for medicaid recipients with depression* (Working Paper). New York: MDRC.

King, G., Gakidou, E., Ravishankar, N. et al. (2007). A "politically robust" experimental design for public policy evaluation with application to the Mexican universal health insurance program. *Journal of Policy Analysis and Management, 26*(3), 479-509.

Kling, J. & Liebman, J. (2005). *Experimental analysis of neighborhood effects on youth.* (NBER Working Paper 11577). Cambridge, MA: National Bureau of Economic Research.

Kremer, M. & Miguel, E. (2007). The illusion of sustainability. *Quarterly Journal of Economics, 122*(3), 1007-1065.

Kremer, M., Miguel, E., Mullainathan, S., Null, C. & Zwane, A. P. (2009a). Making water safe: Price, persuasion, peers, promoters, or product design? Unpublished Working Paper. Harvard University.

Kremer, M., Miguel, E. & Thornton, R. (2009b). Incentives to learn. *Review of Economics and Statistics, 91*(3), 437-456.

Kremer, M., Moulin, S. & Namanyu, R. (2003). *Decentralization: A cautionary tale* (Paper No. 10). Cambridge, MA: Poverty Action Lab.

Kremer, M. & Vermeersch, C. (2004). *School meals, educational attainment, and school competition: Evidence from a randomized evaluation* (World Bank Policy Research Working Paper 2523). Washington, DC: World Bank.

Lalive, R. & Cattaneo, M. A. (2009). Social interactions and schooling decisions. *Review of Economics and Statistics, 91*(3), 457-477.

Liebman, J., Katz, L. & J., K. (2004). *Beyond treatment effects: Estimating the relationship between neighborhood poverty and individual outcomes in the MTO experiment* (NBER Working Paper 493). Cambridge, MA: National Bureau of Economic Research.

Ludwig, J., Kling, J. R. & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives, 25*(3), 17-38.

McKenzie, D., Gibson, J. & Stillman, S. (2010). How important is selection? Experimental vs. Non-experimental measures of the income gains from migration. *Journal of the European Economic Association, 8*(4), 913-945.

McKenzie, D. & Yang, D. (2010). *Experimental approaches in migration studies* (World Bank Policy Research Working Paper No. 5395). Washington, DC: World Bank.

Miguel, E. & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72*(1), 159-217.

Moehler, D. C. (2010). Democracy, governance, and randomized development assistance. *Annals of the American Academy of Political and Social Science, 628*(1), 30-46.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Olken, B. A. (2010). Direct democracy and local public goods: Evidence from a field experiment in Indonesia. *American Political Science Review, 104*(2), 243-267.

Population Services International. (2003). What is social marketing? Washington, DC: Population Services International.

Przeworski, A. & Limongi, F. (1997). Modernization: Theories and facts. *World Politics, 49*(2), 155-183.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5*(4), 472-480.

Schneider, T. R., Salovey, P., Pallonen, U., Mundorf, N., Smith, N. F. & Steward, W. T. (2001). Visual and auditory message framing effects on tobacco smoking. *Journal of Applied Social Psychology., 31*(4), 667-682.

Schultz, P. (2004). School subsidies for the poor: Evaluating the Mexican PROGRESA poverty program. *Journal of Development Economics, 74*(1), 199-250.

Simester, D., Hu, Y. U., Brynjolfsson, E. & Anderson, E. T. (2009). Dynamics of retail advertising: Evidence from a field experiment. *Economic Inquiry, 47*(3), 482-499.

Splawa-Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczynch Tom X*, 1-51.

Stillman, S., McKenzie, D. & Gibson, J. (2009). Migration and mental health: Evidence from a natural experiment. *Journal of Health Economics Journal of Health Economics, 28*(3), 677-687.

Thornton, R. (2005). The demand for and impact of HIV testing: Evidence from a field experiment. Unpublished Working Paper. Harvard University.

Waddington, H. & Snilstveit, B. (2009). Effectiveness and sustainability of water, sanitation, and hygiene interventions in combating diarrhoea. *Journal of Development Effectiveness, 1*(3), 295-335.

Wansink, B. (2006). *Mindless eating : Why we eat more than we think*. New York: Bantam Books.

Watson, T. (2006). Public health investments and the infant mortality gap: Evidence from federal sanitation interventions on U.S. Indian reservations. *Journal of Public Economics, 90*(8-9), 1537-1560.

Webster, M. & Sell, J. (2007). *Laboratory experiments in the social sciences*. Boston: Academic Press.

Williams-Piehota, P., Cox, A., Silvera, S. N. et al. (2004). Casting health messages in terms of responsibility for dietary change: Increasing fruit and vegetable consumption. *Journal of Nutrition Education and Behavior, 36*(3), 114-120.

Wolf, P. J., Silverberg, M., Institute of Education Sciences & National Center for Education Evaluations and Regional Assistance. (2010). *Evaluation of the DC opportunity scholarship program: Final report*. Washington, DC: Institute of Education Sciences and National Center for Education Evaluation and Regional Assistance.

World Health Organization. (August 16, 2007). Who releases new guidance on insecticide-treated mosquito nets. *World Health Organization News Release*. Retrieved from www.who.int/mediacentre/news/releases/2007/pr43/en/index.html

---

Please cite this document as:

Gerber, A.S., Green, D.P., and Sovey, A. (2012). Evaluating law and policy using randomized experiments. *PHLR Methods Monograph Series*.