A "Politically Robust" Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program

A "Politically Robust" Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program

Gary King
Emmanuela Gakidou
Nirmala Ravishankar
Ryan T. Moore
Jason Lakin
Manett Vargas
Martha María Téllez-Rojo
Juan Eugenio Hernández Ávila
Mauricio Hernández Ávila
Héctor Hernández Llamas

Abstract

We develop an approach to conducting large-scale randomized public policy experiments intended to be more robust to the political interventions that have ruined some or all parts of many similar previous efforts. Our proposed design is insulated from selection bias in some circumstances even if we lose observations; our inferences can still be unbiased even if politics disrupts any two of the three steps in our analytical procedures; and other empirical checks are available to validate the overall design. We illustrate with a design and empirical validation of an evaluation of the Mexican Seguro Popular de Salud (Universal Health Insurance) program we are conducting. Seguro Popular, which is intended to grow to provide medical care, drugs, preventative services, and financial health protection to the 50 million Mexicans without health insurance, is one of the largest health reforms of any country in the last two decades. The evaluation is also large scale, constituting one of the largest policy experiments to date and what may be the largest randomized health policy experiment ever. © 2007 by the Association for Public Policy Analysis and Management

INTRODUCTION

The history of public policy experiments is littered with evaluations torpedoed by politicians appropriately attentive to the short-term desires of their constituents, such as those who wind up in control groups without new services or who cannot imagine why a government would randomly assign citizens to government programs. The fact that a scientific evaluation might maximize the interests of people in the long run is often no match for the understandable outrage of constituents and the embarrassment politicians may suffer in the short run. Scholars need to remember, however, that responsive political behavior by political elites is an integral and essential feature of democratic political systems and should not be treated with disdain or as an inconvenience. Instead, the reality of democratic politics needs to

Journal of Policy Analysis and Management, Vol. 26, No. 3, 479–506 (2007)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20279

be built into evaluation designs from the start, or else researchers risk their plans being doomed to an unpleasant demise.

Thus, although not always fully recognized, all public policy evaluations, including ours, are projects in both *political* science and political *science*. We try to account for this issue explicitly by developing a general randomized design that has features which should enable an evaluation to survive even if certain portions of it are destroyed through unexpected or ill-timed political interventions. Although most of the individual features of our design have been used in prior research, their advantages in accommodating political realities have only rarely been recognized and chosen for this purpose—especially for experiments in the developing world. These features may also be of use to other researchers designing policy research in these necessarily political environments.[1]

We also report on applying our design to a large-scale evaluation of Seguro Popular de Salud (SPS) we are conducting. SPS is a program of the Mexican federal government designed to extend medical services, preventive care, pharmaceuticals, and financial health protection to the approximately half of the Mexican population that had no regular access to health care, particularly those with low incomes. In terms of the national geographic coverage, the substantial cost of the program, the extent of the benefits available to individuals, or the "aim to provide social protection in health to the 50 million uninsured Mexicans" (Frenk, Sepúlveda, Gómez-Dantés, & Knaul, 2003, p. 1667), SPS represents one of the largest health policy reforms in the world in the last two decades. SPS is highly visible and politically sensitive, and was a prominent issue in the 2006 national election. In addition, because of the importance of the evaluation to the Mexican government and the many politicians at every level of government who could influence the program or evaluation—from the leaders of the federal government, to the state governors, to national and state legislators, to SPS program administrators at the federal and state level, and so on, all the way down to administrators of local health care clinics and even frontline care givers—we may even be especially vulnerable to the side effects of enterprising politicians attempting to please their constituents. As such, although we believe that the randomized evaluation design we propose here may find more general applicability, it may be especially valuable in contexts like the SPS evaluation.

We first give some examples of political and other factors that affected previous large-scale experiments and then offer a brief overview of the SPS program and the origins of this evaluation. We then describe our experimental design, the expected effects of SPS, and an empirical validation. The appendices briefly list variables available in our survey and describe our statistical analysis plans for the post-experimental treatment period.[2]

LESSONS FROM EXPERIMENTAL FAILURES

"Evaluation often confronts awkward political issues" and may even impose "personal costs to public servants" (Lewis, 2005, p. 202). Experiments conducted in ongoing public policy programs, like ours, may have advantages in realism and external validity, but they also pose special problems due to constraints imposed by politicians and program administrators, and their interactions with subject expectations,

[1] The concept of a research design that survives even if randomization does not is occasionally mentioned in the literature, such as the Cook and Campbell (1979, p. 134) concept of "fallback," which is sometimes implemented via before–after designs or some matching strategies (Flay & Best, 1982).

[2] All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

substitution effects, and the correspondence between the experimental treatment and actual program benefits (Burtless, 1995; Heckman, 1992). In fact, two of the three conclusions of the participants in a European Regional Consultation (including World Health Organization [WHO], several health ministries, Organisation for Economic Co-operation and Development [OECD], and the World Bank) about how to improve the analysis of health policy implementations were more consideration of local "political context" and speaking to "the political concerns of policymakers more explicitly" (Murray & Evans, 2003, p. 61). Not only are experiments increasingly becoming an important part of political science (Green & Gerber, 2002), but experimenters, particularly those evaluating public policies, also need political science. This may be particularly true in the developing world.

For example, immediately before the start of one of the treatment periods in the evaluation of the Mexican Programa de Educación, Salud y Alimentación (PROGRESA) antipoverty program, now called Oportunidades (Adato Coady, & Ruel, 2000; Gertler, 2000, 2006), people in control areas who would not receive services located adjacent to treatment areas, along with program administrators and state governors, convinced federal administrators in charge of the program to include them in the treatment group so they could also receive services (Greenberg & Shroder, 2004, p. 436). Although this was good news in the short run for these constituents, it potentially biased an aspect of the experiment. In fact, results from early stages of the evaluation that demonstrated the program's benefits led to considerable popular pressure to end the evaluation and give services to everyone.

This was hardly a unique, or even unusual, occurrence in field experiments. In Project STAR, a large education experiment designed to test the effects of class size, about 10 percent of the students were moved to classes of different sizes than ones to which they were randomly assigned at first, in part because of parental complaints and organized lobbying (Krueger, 1999; Dee & Keys, 2004). In a subsidized meal program in Kenya, upset parents in over half of the control schools organized to raise funds for student meals to match what was being received in the treatment group (Greenberg & Shroder, 2004, p. 399). In 1980, a field experiment conducted for the local government in Stockholm was to consider expanding a bus route to a major hospital and factory, but because at the last minute the trade unions objected to the experiment, almost no subjects showed up (Bohm, 1984; Harrison & List, 2004).

Heckman and Smith (1995, p. 100) point out that over 90 percent of administrators of training centers approached for the U.S. Department of Labor's Job Training Partnership Act evaluation refused participation in the experiment, the most commonly given reason for which was "ethical and public relations concerns." These officials were presumably worried about negative results casting doubts on their program or their own performance. Administrators of a Labor Market Training study in Norway circumvented randomization by selectively declaring enough subjects ineligible so that the remaining subjects numbered only one more than the number of available training slots to which they were to be randomly assigned (Torp, Rauum, Hernaes, & Goldstein, 1993). In a British job training experiment called Job-Plan, more than 20 percent of controls received the training workshop treatment merely because they asked to participate or were mistakenly required to do so (Greenberg & Shroder, 2004, pp. 446–447). A governmental agency, the California Youth Authority, abruptly stopped conducting randomized experiments altogether for direct "political and related ideological pressures" (Palmer & Petrosino, 2003). Indeed, "the potential list of problems is endless" (Nickerson, 2005, p. 283).

Of course, researchers in many of the experiments described in this section found ways to contribute valuable information about their intended subjects, despite their

difficulties. And issues that are not as essentially political can also threaten research designs, such as "the incongruence between treatment assignment and receipt" (Camasso et al., 2003), complicated self-selection issues (Howell, 2004), and sample attrition (Greenberg et al., 2006). But as Boruch (1997, pp. 182–184) writes, "Judgments about the capacity of a site to engage in a controlled field test at times require dedicating serious attention to the site's political environment. [In evaluations,] the possibility of failure is real. It must be planned for."

THE SEGURO POPULAR POLICY INNOVATION

The plan the Mexican government passed began with a pilot phase in 2002 in 5 states, and by the end of the first year was present in 20 states with about 296,000 families affiliated. The law that formally created SPS as part of the "Sistema de Protección Social en Salud" (System for Social Protection in Health) was a 2003 and 2004 modification of the Mexican General Health Law. Under the law, the Comisión Nacional de Protección Social en Salud (National Commission for Social Protection in Health) is in charge of supervising the system as a whole and coordinating with the state offices of the program. Although the literal translation of "Seguro Popular" is popular or universal insurance, and protection from the impoverishment that can result from catastrophic health expenditures is a primary aim, it is not intended to be a self-sustaining insurance program, and indeed the Spanish word for "insurance" does not appear in the authorizing legislation. SPS is instead a social welfare program that provides preventive and regular health care, as well as subsidies to reduce out-of-pocket health expenditures, primarily to lower income uninsured Mexicans, and aims to strengthen the certification and effectiveness of local health facilities.

The federal government spent the equivalent of $795.5 million on SPS in 2005, which was entirely new money spent on the health sector. When fully implemented, they intend for SPS to increase total health spending in the country by an additional 1 percent of GDP compared to 2002.

As SPS operates now (August 2006), individuals must formally affiliate with SPS to receive medical care. When they affiliate, SPS covers 249 health interventions outlined by the Universal Catalog of Health Services, including the provision of 307 drugs associated with the services. These interventions treat the diseases responsible for about 95 percent of the burden of disease in Mexico. Affiliated families pay a semiannual or annual quota that increases by decile of income, with the lowest two deciles exempt. The largest share of the lowest two deciles are enrolled in the Oportunidades antipoverty program and are formally affiliated with SPS automatically when an area is ready to enter the program. (To access services requires both formal affiliation and individual knowledge of this status, and so we will also see how much of an advantage automatic rather than self-affiliation turns out to be.) The federal government provides a contribution to the states for each family affiliated, supplemented by a social quota per family from the states. Each year, an office of the Health Ministry, independent of the National Commission, certifies only those communities that have adequate medical facilities and decides, in part on that basis, on the number of families each state is funded to attempt to affiliate. Readying areas for affiliation thus requires state contributions as well.

The program is being rolled out in stages, increasing coverage each year. By the end of 2003, 24 states were participating, with 614,000 families affiliated, and by 2005, all 32 states had some areas included, with 3.5 million families affiliated. The entire uninsured population is expected to have the opportunity to affiliate by 2010, but, because they would have to pay for SPS services and can choose to receive medical

services elsewhere, we expect that many households in higher income deciles will not find it attractive enough to affiliate.[3]

SPS represents a large part of a massive reform and constitutes one of the main policy changes of the Fox administration. Passing the reform itself was an unexpected outcome of divided government in Mexico, advantaged by the support of most of the governors (Lakin, 2005). Only the Distrito Federal (Mexico City) did not participate in SPS when we started our evaluation, although it is included now. The mayor of Mexico City, who would later announce as a candidate for president, implemented his own competing health program and was not a supporter of SPS.

ORIGINS OF THE EVALUATION

Although a constitutional term limit means that the Mexican government that promulgated the plan could hold office for only one six-year term, those who designed SPS intended to create a permanent entitlement that lasts well beyond the current government. How one democratically elected government can "tie the hands" of, or even influence, their democratically elected successors is a fundamental question of practical governance as well as of normative democratic theory (Klarman, 1997; Posner and Vermeule, 2002; Sterk, 2003). Although formal "entrenching legislative rules" are often illegal, any change in the status quo can build citizen expectations, alter international commitments, change the division of legislative votes needed to pass alternative legislation, and otherwise constrain the choices of future governments. Scholars have developed formal theories (Alesina & Tabellini, 1990), extended case studies of specific entrenched policies (Derthick, 1979), systematic empirical evidence (Franzese, 2002), and philosophical arguments (Thompson, 2005) that elaborate on the consequences of this crucial commitment problem.

Mexican President Vicente Fox Quesada and Health Minister Julio Frenk Mora presumably had strategies like these in mind, but they also implemented an open plan for the scientific evaluation and persistence of their program. Their Ministry of Health (MoH), and the independent National Institute of Public Health (INSP), commissioned the Harvard University Team among the authors of the present paper to lead an independent, ongoing scientific evaluation of SPS. Their theory was that if we concluded that the program is a success, the next government would be less likely to want to eliminate it and might even find it more difficult to do so even if they wanted to. The benefit to the government, just as in science, is greatest when the hypothesis is most vulnerable to being proven wrong. And they accordingly have made themselves highly vulnerable because, if SPS or some portion of it fails, we will say so as clearly as we will if it succeeds. We do not know whether this justification will work in other evaluations, but it seems to have worked here and to be a reasonable hypothesis that it might work in other situations.

[3] The statistics in this section appear in reports available at http://www.seguro-popular.gob.mx/. An English translation of Article 8 (Transitory) of the law states, "From the date that this Decree takes effect, every year and in a cumulative manner, of those families who are eligible for new incorporation, up to 14.3 percent will be able to become incorporated into the System for Social protection in Health, with the objective of achieving a 100 percent coverage by the year 2010. In the fiscal year 2004 and subsequent years, families could be added, whose incorporation could be paid with resources from the Health Services Contributions Fund to which the Fiscal Coordination Law makes reference, resources for programs of Administrative Chapter 12 Health of the Federal Budget, and resources for the function of health, requested by the federal government. for the System for Social Protection in Health and approved by the Chamber of Deputies. Coverage of the services for social protection in health will start by giving preference to the population in the first two income deciles in areas with greatest deprivation, rural and indigenous areas, in compliance with the registries kept by the federal government." See also Frenk, Gómez-Dantés, Lezana, & Knaul (2006).

The MoH provided us access to government officials and experts on SPS, information on the inner workings of the program, the ability to influence how SPS was implemented so that we could more easily and rigorously evaluate it, and the means to design and direct data collection plans. The officials requested no prepublication approval of our conclusions.

Of course, like any public policy program, some parts of SPS will likely work and others will probably not perform as expected. Thus, the main purpose of our ongoing evaluation will probably not be a dichotomous declaration of victory or defeat for the hypothesis that SPS succeeded, but rather a process of using modern tools of social science to learn about how to improve the program and ultimately the health of the Mexican population (see Heckman & Smith, 1995, p. 94).

EXPERIMENTAL DESIGN

We now describe our experimental design, detail the political and other issues that arose in developing and then implementing it, and explain the choices and solutions we made along the way. Briefly:

1. We define 12,284 contiguous geographic regions that tile Mexico's 31 states. We call these "health clusters," each one of which includes an actual or future health clinic or facility and the population catchment area around it.
2. We persuaded 13 of the 31 Mexican states, to participate in the evaluation, which was composed of 7,078 (5,439 rural and 1,639 urban) health clusters.
3. We matched these health clusters in pairs so that members of each pair were as similar as possible on a range of background characteristics.
4. For the first cohort of our experimental study, we selected 74 of these pairs of health clusters from 7 states, portrayed in Figure 1, with selection based on closeness of the match, likelihood of compliance with the experiment, and necessary political and other criteria. (These 148 health clusters include 1,380 localities, approximately 118,569 households, and about 534,457 individuals. We expect subsequent experimental cohorts, which we are now selecting, to be roughly the same size.)
5. We randomly assigned one health cluster from each pair to receive encouragement to individuals to affiliate with SPS, along with the health facilities, drugs, and doctors necessary to implement the program effectively. The other health cluster in each pair received nothing extra.
6. At the time of random assignment, we conducted a baseline survey of the health facility within each health cluster, and a survey of about 32,000 randomly selected households within 50 of the 74 pairs of clusters (chosen based on likelihood of compliance with encouragement to affiliate and similarity of the clusters within each pair). We used this baseline household survey to verify that the treated and control groups are similar on a wide range of health characteristics and other variables. (We do not analyze the health facilities survey in this paper.)
7. Ten months after random assignment, and then repeatedly at other intervals, we conducted follow-up surveys of the health facilities and individuals within each health cluster, which we used to ascertain the effect of the program.

We now discuss our dependent variables and the surveys we are fielding to measure them, how we found and defined a politically acceptable level at which to randomize, how we insulate ourselves from selection bias in some circumstances even if political interventions cause us to lose some of our observations, and the

Figure 1. Mexican states participating in the first evaluation cohort.

triple robustness property of our evaluation design and analysis strategy. We then discuss limitations of our design.

Survey Measures of Program Outcomes

A public policy program like SPS has many targeted goals and multiple measurable intermediate milestones along the way necessary to achieve the goals. We are collecting data on our outcome measures via specially designed surveys of approximately 32,000 individuals and a separate survey of the staffing and conditions at the health facilities, both within the health clusters selected for our experiment. The individual-level survey involved a random probability sample of households in these areas, an interview with one person in each household who knew the most about the household and its members, and one additional randomly selected individual over age 18 (weighted via Kish tables to be representative; see Kish, 1949).

The variables measured include satisfaction with the health care provider, health self-assessments, self-assessments of chronic conditions, and reports of risk factors and health conditions (a detailed list appears in the Appendix).[4] In addition to the traditional survey items, we also include physical testing of blood pressure, cholesterol, blood sugar, and HbA1c, the last two being indicators of diabetes. For many respondents, having the medical tests and being offered immediate results were a great motivation to participate in our study. Paradoxically, from the perspective of surveys in the U.S., we greatly reduced nonresponse problems by telling respondents about the medical tests at the outset and administering them—including three separate finger pricks to draw blood—only as the very last step in the survey.

[4] All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

The survey also included the standard battery of sociodemographic variables, as well as proxies for several political variables to control for the possibility of respondents biasing self-assessment answers to favor their preferred party in the 2006 election. The political variables would ideally have included party identification and voting behavior questions, but we were advised (by the Federal Electoral Institute) to avoid these so close to the election. We thus included reasonable proxies for these in questions asked about whether the government should try to reduce differences between rich and poor, whether the electricity industry should be privatized, and whether government spending was aimed at the needs of the citizens.

We have already conducted one baseline survey, which was fielded at the time of randomization of clusters to treated and control groups, in August 2005. On the advantages of baseline surveys on improving precision, see Bloom, Richburg-Hayes, and Black (2007); Glazerman, Myer, and Decker (2006). In addition, the true causal effect of treatment on the outcome measures in this survey is zero, because treatment was assigned at almost the same time and could not have had any real effect yet. In the Empirical Validation section below, we estimate these causal effects on the baseline, and use the frequency of estimates that deviate from zero as an indication of potential problems with our evaluation design. We plan a repeat survey of the same respondents approximately 10 months later to see the early effects of the program, and then several other surveys at longer intervals.

Politically Acceptable Randomization

Experimentation is best conducted via (1) random selection of subjects from the population of interest, (2) random assignment of these subjects to treated and control groups, and (3) a large number of subjects. Random selection is typically infeasible in large scale policy experiments (except within local, nonrandomly selected areas), and was infeasible for aspects of our experiment too. Studies without random selection that are otherwise methodologically sound can produce valid causal inferences for the subjects of the experiment, although not necessarily for the population at large. The practice in medical research, where random selection is rarely feasible, is to repeat such experiments in many areas with diverse subject pools until generalization to the larger target population becomes more plausible. Although the expense of large scale field experiments often make repeating the experiment difficult, we plan the same strategy in our evaluation by using multiple waves of subjects in cohorts selected at different times throughout the country. (Indeed, we have already begun the selection of our second experimental cohort.)

In contrast to random selection, random assignment of values of the treatment variable is normally considered the sine qua non of experimental design. Its importance stems from guaranteeing in large samples that the treatment is unrelated to any potential confounding variables even if those variables are not observed or known. Randomization of medical care to individuals would avoid problems such as sick citizens signing up to receive care more than the healthy, which might cause one to conclude that the program made people ill even if the reverse were true. Unfortunately, individual-level randomization is often politically unacceptable in policy evaluations because government benefits are designed to be withheld only for some reason having to do with an individual's qualifications. Researchers often think of randomness as the ultimate in fairness, because the assignments it produces would be the same even if the name of the person receiving the random number changed; but this anonymity property is often viewed by citizens and politicians as the ultimate in whimsy, ignoring as it does how much the person not receiving the services needs them. In fact, in

medical research, potential subjects are less likely to agree to participate in experiments if randomness is part of the design (Kramer & Shapiro, 1984).

Random assignment was especially suspect in the Mexican government, given the political problems accompanying it in the last large scale policy evaluation (of Oportunidades, described above), and researchers from INSP and MoH first told us that random assignment was impossible. However, program implementation always includes some arbitrary decisions, normally made by lower level administrative officials without the attention of political elites. When decisions are recognized as arbitrary, randomizing those decisions becomes acceptable. Because some decisions are always made below the level of political radar, we offer the generalization that *randomization is always acceptable at one level below that at which politicians care*. Once the right officials understood this point, it was easy to search together with them to find the most informative way to randomize subject to reasonable political constraints (see also Green & Gerber, 2002, p. 821).

In the SPS evaluation, we could not randomize individuals to affiliation because it would have been politically and ethically unacceptable, but also because every citizen is technically permitted to affiliate even when no health facility is nearby. The level of random assignment we chose is the *health cluster*, which is a geographic unit we defined for the purposes of the evaluation. We define a health cluster as an actual or planned health clinic (Clínicas, centros de salud, hospitales, etc.), and the catchment area around it. Put differently, one can never randomize entitlements, and SPS is designed as an entitlement at the individual level. However, SPS is not an entitlement at the cluster level while the program is being rolled out. Our study is an example of what is variously called a "place-randomized" (Boruch et al., 2004), "group-randomized" (Murray, 1998), or "cluster-randomized" (Donner & Klar, 2000) trial.

To construct these clusters, we worked with the *Núcleo de Acopio y Análisis de Información en Salud* at INSP and, in negotiation with the state governments, first mapped the location of every current or planned health clinic in the country and then attempted to define the catchment area around each as travel time of less than one day to the clinic. For "travel time" we used graphic information system (GIS) technology to approximate the actual time it takes for an individual in each household, using transportation methods available, to travel to the closest health clinic where he or she could receive care (rather than the linear distance "as the crow flies"). We attempted to account for factors such as available roads; whether the roads were used for cars, public transportation, or walking; and natural boundaries like rivers without bridges. We used localidades (localities) in rural areas as building blocks, but within AGEBs (Area Geoestadística Básica, which correspond roughly to U.S. census tracts) in urban areas, we used detailed street-level information and location of the health facility. With helpful checks performed by the states, we defined 10,616 rural and 1,668 urban health clusters nationwide, and together these 12,284 clusters tile the whole country, other than Mexico City, which did not participate in SPS at the time.

Health cluster-level random assignment was politically feasible because, even without our experiment, SPS must be rolled out to different parts of the country over time. This is the case because funds, health clinics, doctors, and drugs do not exist to give everyone access all at once, and so affiliation will have to be explicitly encouraged in some areas, and other areas will need to wait. The special advantage of health clusters as the unit of randomization in this context is that it is effectively the level at which the policy decision to roll out the program is made, the level at which funds are spent, and the level at which health clinics are located, built, stocked, funded, and staffed. It is thus both administratively feasible and enables us

to estimate at least one causal effect, in these or similar areas, at the level of interest and of most relevance to policymakers who would choose to implement or roll out the program to new areas.

In addition to the causal effect on the policy decision to implement the program in a health cluster on the health and well-being of its population, we would also like to estimate the effect of any one individual's affiliation with SPS on that person. Although we did not randomly assign affiliation at the individual level, we can use the random assignment of health clusters (in what is called an "encouragement design") to estimate the causal effect of individual level affiliation, as if it were randomly assigned.[5]

The particular health clusters to be randomized must be chosen from those ready for affiliation and politically feasible to randomize. We started with all 12,284 health clusters and then eliminated areas from the experiment in five categories. First are areas that the state governors and their administrations decided should receive SPS no matter what, and thus are not subject to our experimental assignment; these decisions may be for whatever technical, policy, or political reasons the officials deem appropriate. We were not able to conduct a detailed study of how these decisions were made, as they were the result of a complicated negotiation process between the SPS administration and the states. Second, we eliminated areas in states that were not yet participating in SPS. Third, because providing the financial means to use health care is useless when doctors, hospitals, or drugs are unavailable, areas with inadequate or nonexistent health facilities that the government could not improve in our time frame were excluded from both SPS and the experiment, at this stage. Fourth, we eliminated from the experiment areas with which many families were affiliated prior to our experiment, because random assignment would have had little effect (or in other words, our encouragement to affiliate would likely be ignored). And finally, we dropped very small rural clusters (under 1,000 population) and kept only those urban clusters with more than 2,500 and fewer than 15,000 population.

Then, during the annual negotiation between the states and the federal government on which areas will receive the go-ahead to begin affiliating families with SPS, we were offered a large number of health clusters we could randomize. The largest number we could afford to collect data on was 148, which we chose to optimize our matching criteria and compliance with the experiment (we describe these procedures in a separate section below). This strategy was politically acceptable because the original plan for SPS was to phase it in over six years, and so we are able to exploit the natural phase-in delay in the program to encourage affiliation in randomly selected treated areas and to do nothing in control areas. All clusters would eventually be included in SPS and no absolute restriction was placed on individual affiliation at any time. Our baseline and follow-up surveys are conducted within these clusters.

[5] The basic idea of an encouragement design is to use health-cluster random assignments as an instrument with known properties to estimate the direct effect of affiliation (for example, Hirano, Imbens, Rubin, & Zhou, 2000; Frangakis, Rubin, & Zhou, 2002; Barnard, Frangakis, Hill, & Rubin, 2003). The key issue is ascertaining who complies with the experiment—affiliates when encouraged to and does not affiliate when not encouraged—which can be estimated directly with this design. In most cases, we expect few individuals to affiliate and use services in areas not encouraged by our experimental assignment, because they would need to travel far to affiliate, and then when affiliation takes effect 30 days later, would have to travel back for any needed medical treatment. Lower income individuals in randomly assigned encouragement areas are highly likely to sign up for the program, as it is free or inexpensive for them, whereas upper income people who have their own health insurance and separate hospitals are much less likely to affiliate. Oportunidades families, which constitute roughly 90 percent of families in the lowest two deciles of income, are affiliated by the government automatically.

We measure outcome variables at the level of the health facility for all 148 clusters and (due to financial constraints) at the household and individual level for 100 of these (selected from the 148 with rules we describe below). Although we describe what we plan to do with both, we only analyze the individual-level baseline survey in this paper.

Losing Clusters without Losing Balance

The most common experimental design is *classical randomization*, which in our application would assign each health cluster to the treated or control group based on a separate coin flip. This design makes it possible to base inferences on a simple difference in means between the two groups, because the observed and unobserved characteristics of the control and treated clusters are the same, at least on average. Randomization, then, makes it possible to avoid resorting to the usual model-dependent regression adjustments that are required in observational studies. Classical randomization works fine if all health clusters in the study at the start remain in until the end. However, if even one cluster is lost—due to political intervention, measurement errors, incorrect randomization, or for any other reason—we would then no longer be guaranteed that the treated and control groups are comparable on average, and the benefits of randomization would be lost.

Any loss of observations in a classical randomization study can thus result in imbalance between the groups, which can generate bias. For example, the PROGRESA evaluation described above used classical randomization and had some loss of observations. Although empirical evidence in that study did "not indicate any systematic differences" between the treated and control groups on the observed variables, the randomization no longer guarantees that any unobserved variables must be similarly balanced on average across the groups (Behrman & Todd, 1999, p. 8).

Especially given this previous experience, we must expect to lose health clusters, and so we need a design that allows some clusters to be lost, under at least some circumstances, without also losing the advantages of randomization. Thus, we turn to what is known as a *randomized cluster matched pair design*, which, if used appropriately, has a self-protecting property that has rarely been discussed in print, even though it may have been used in practice (Donner & Klar, 2000). In matched pair randomization, we first select pairs of health clusters that are matched, or at least as similar as possible, on a large set of available background characteristics. Then, by flipping a coin, we randomly choose one of the two clusters within each pair to receive treatment and the other to be the control. The result of this process is exact balance between the entire treated and control groups of health clusters on all variables included in the matching for which exact matches among the clusters are available, or near matches otherwise. Variables not matched on are balanced by randomization and therefore only match on average.

Matching on covariates before randomization in this way (compared to classical randomization) "can increase balance on these covariates, increase the efficiency of estimation and the power of hypothesis tests, and reduce the required sample size for fixed precision or power," and if the covariates are unrelated to variables in our analysis, matching "does not harm statistical efficiency or power" (Greevy, Lu, Silver, & Rosenbaum, 2004, p. 264). Matching before randomization thus does not seem to have significant disadvantages, except in much smaller sample sizes than we have, where efficiency is still improved (Imai, King, & Stuart, 2007), but power can be reduced (Klar & Donner, 1997; Raudenbush, Martinez, & Spybrook, 2007); it also possesses other advantages discussed below.

The key additional advantage of the matched pair design from our perspective is that it enables us to protect ourselves, to a degree, from selection bias that could otherwise occur with the loss of clusters. In particular, if we lose a cluster for a reason related to one or more of the variables we matched on, such as low-income areas or clusters within cities, then no bias would be induced for the remaining clusters. That is, whether we delete or impute the remaining member of the pair that suffered a loss of a cluster under these circumstances, the set of all remaining pairs in the study would still be as balanced—matched on observed background characteristics and randomized within pairs—as the original full data set. Thus, any variable we can measure and match on when creating pairs removes a potential for selection bias if later on we lose a cluster due to a reason related to that variable. Selection bias might still occur under this design if, for political or other reasons, clusters were lost after the start of the study for reasons both unrelated to our matched variables and related to the treatment assignment, or by selecting on the casual effect, but we would be fully protected from bias due to any variable we were able to match on. Classical randomization, which does not match on any variables, lacks this protective property.[6]

A Triply Robust Evaluation Design

A key part of our evaluation design includes (1) paired matching of health clusters, (2) randomization of treatment and control within pairs, and (3) parametric adjustment to estimate the quantities of interest, each of which we describe in this section. Under weak regularity conditions, when any one of these steps works as planned, we will be able to make valid causal inferences even if the other two parts fail. We call this property *triple robustness* (see Robins & Rotnitzky, 2001; Ho, Imai, King, & Stuart, 2007).[7]

An Algorithm for Paired Matching

The most commonly used matching algorithms are designed to apply to data for which the treatment assignment is known prior to matching (Ho et al., 2007). In our problem, which is known mathematically as "nonbipartite matching" and creates n pairs from $2n$ health clusters, pairing must be completed prior to treatment assignment. Optimal algorithms have been developed for this problem that are appropriate when all clusters are randomized simultaneously (Greevy et al., 2004).

In our evaluation, however, only a simplified textbook-like summary of our procedures would sound like we had *simultaneously* randomized all our clusters.

[6] Randomized matched pair designs also have the advantage that they can be used to provide pair-level causal effect estimates. Indeed, a noisy estimate would be, for any outcome measure, the treated value minus the control value in the same pair. Statistical techniques can also be used to reduce the noise. Pair-level causal effects can provide valuable information if SPS is having different effects in different parts of the country, or is more or less successful for certain types of population groups. For example, we suspect that SPS will have a bigger impact in low-income, rural areas, because those are the areas for which it was primarily designed. Other possibilities could also be explored but would, of course, remain more uncertain and in need of replication in other cohorts.

[7] Each of the three components of our design have been used before separately, and sometimes in combinations, in previous research. Paired matching is a special case of "blocking" in the experimental design literature, where the general advice has long been to "block what you can and randomize what you cannot" (Box, Hunger, & Hunter, 1978, p. 103). Paired matching typically provides higher levels of variance control than other forms of blocking. To our knowledge, the triple robustness property has not been noted directly before, nor have its advantages for creating fail-safe experimental research designs.

Unfortunately, conducting an experiment in the real world of politics and policy is not remotely as controlled as most textbook discussions of research design. We constructed our matches in real time, while the SPS program was being rolled out, under conditions of uncertainty and considerable time pressure. At the same time, the states and the federal government were negotiating on which regions had health care facilities above the threshold for qualification, how much money would be available in this round for affiliation, and which health clusters would be in the experiment and thus subject to our decision about who would get SPS. During this time, information on the geographic location, and thus definition of individual health clusters, was improving, data coding background covariates were being corrected, and our data sets were being continually updated. Simultaneous matching was also not desirable, because we could only afford to conduct our individual-level survey in a subset of the randomized pairs, and so we wanted to optimize better with respect to this subset than we could with simultaneous matching in a larger group.

There were also inevitable misunderstandings along the way, such as when an early attempt at randomization caused some states to inform us that we should discard two-thirds of the pairs we thought we had randomized. Upon investigation, we found that the states wanted to allow only the pairs in the experiment where the cluster assigned treatment was the one (of the two) they wanted to receive SPS. We explained that investigator control of the experiment was essential for scientific randomization (a procedure that introduces no bias because randomizations are by definition mutually independent), and so we began the process from scratch.

Because optimal matching of the entire set of clusters all at once was both infeasible and undesirable, we designed a new algorithm better suited to the political problems we faced. We call this an *optimally greedy* algorithm. Whereas *optimal* algorithms simultaneously adjust all pairs to optimize a global objective function (such as minimizing the average distance between members of each pair), classic *greedy* algorithms find the closest match for each cluster one at a time. Greedy algorithms are not invariant to the order of matching, and typically match in arbitrary order, such as by observation number, but have the advantage of finding the best match for any one cluster among those available to match. In contrast, our optimally greedy algorithm minimizes the minimum distance between clusters within pairs across the entire set of data available at any one time to match. The arbitrariness of greedy matching is thus avoided, and the advantages of optimal matching are available for any one set of clusters considered together. This algorithm also met our needs because we would only be able to conduct our individual level survey in some, but not all, of our clusters, and so wanted to use the best matches there but still use the full set of pairs to analyze the facilities survey, which would be fielded in all the pairs.

To apply any matching algorithm requires a metric to measure the distance between the clusters within each pair. In our case, we exact match on state and urbanicity, and within those strata use the Mahalanobis metric to compute distances. The Mahalanobis distance is a measure of the difference between the values of all the control variables in the treated and control clusters.[8] The complete procedure, then, is as follows: within a state-urbanicity stratum, compute the Mahalanobis distance for every possible pair of clusters available to be matched at any one time; choose the

[8] In computing the Mahalanobis distance, all the different variables are normalized to the same scale via the variance matrix computed from the observed data to be matched. To reduce sensitivity to outliers in small samples, we improve on this procedure by estimating this matrix from the largest set of health clusters available to us at the time of matching.

pair with the smallest distance and remove it; and repeat until all clusters in the stratum are matched.

Ideally, the background characteristics would include the outcome measures we wished to study, measured prior to randomization, but these were not available. Instead, we followed the usual procedure and included all available plausibly relevant background characteristics. These variables represent demographics distributions, housing infrastructure, insurance, population, characteristics of health facilities, disabled indicators, literacy, geographic characteristics, SPS program participation, income, and others.[9] Although changes in this list led to different pairs being matched, the differences did not seem overly sensitive to specification.

The great advantage of matching geographic areas is that we can always learn more about a pair of clusters than the information our quantitative data indicates by simply visiting the area or talking with those who are familiar with it. When matching individuals in surveys, or other anonymous units, this kind of external qualitative information is typically unavailable, and in fact the particular units matched rarely make an appearance in publications. In our work, we studied geographic maps like those in Figure 2 and researched the pairs found by our algorithm. We used this process mostly to find data errors and to suggest new variables to include in our matching algorithm. Although a similar procedure might cause one to modify the quantitative matches, our discussions with local officials indicated that this did not seem necessary. We found this result somewhat reassuring, that we had matched on all the relevant background characteristics and especially all the ones that the politicians and officials seemed to be immediately aware of.

The success of our matching procedure was largely a function of how many clusters we could persuade officials to make available to us for matching. The more that were available, the better matches we were able to find. To evaluate the quality of the matches, we plotted numerous graphs like that on the left side of Figure 3. This particular graph gives a histogram of the absolute value of the difference in the proportion of the over-65 populations between clusters within each pair. As can be seen for this particular variable, most of the clusters stack up at very nearly zero difference, as we would want, with some others scattered at slightly larger differences. We found similar results for many other graphs of the variables we used to match.

We also used the Mahalanobis distance metric to summarize the differences within the pairs on all variables, an example of which for rural clusters is displayed on the horizontal axis in Figure 4. The horizontal position of the clusters on the graph

[9] The specific variables included in both urban and rural clusters include total population, average education, average rooms per occupant, percent female, percent between 0 and 4 years old, percent under age 18, percent with and without social security rights, percent over 5 years old who were living in the localidad (or AGEB) in both 1995 and 2000, percent disabled, percent married, percent Catholic, percent employed, percent employed in the secondary/tertiary sector, percent living in households making less than twice the minimum wage, percent of households headed by women, weighted marginalization index, a series of housing infrastructure variables (condition of the walls, roof, floor other than dirt, electricity, water access, sewer, other plumbing, and refrigeration), percent in Oportunidades, health infrastructure variables (day beds, consultorios, doctors, and nurses). In addition, we added for the rural clusters percent over 15 years old and illiterate, the percent employed in primary sector, percent over age 5 speaking an indigenous language, and an index comprising the sum of a series of health infrastructure variables describing characteristics of health facilities within 120 minutes of the cluster, area, dummy for affiliation of the health center to SPS, population within 1km without social security. And we added for the urban clusters percent over age 15 and literate; the percent over age 60, and over age 65; percent in IMSS; altitude; and the number of operating rooms, and general or overnight beds, in the nearest medical facility. Of these variables, four had a few missing values, which we multiply imputed as described in King, Honaker, Joseph, and Sheve (2001).

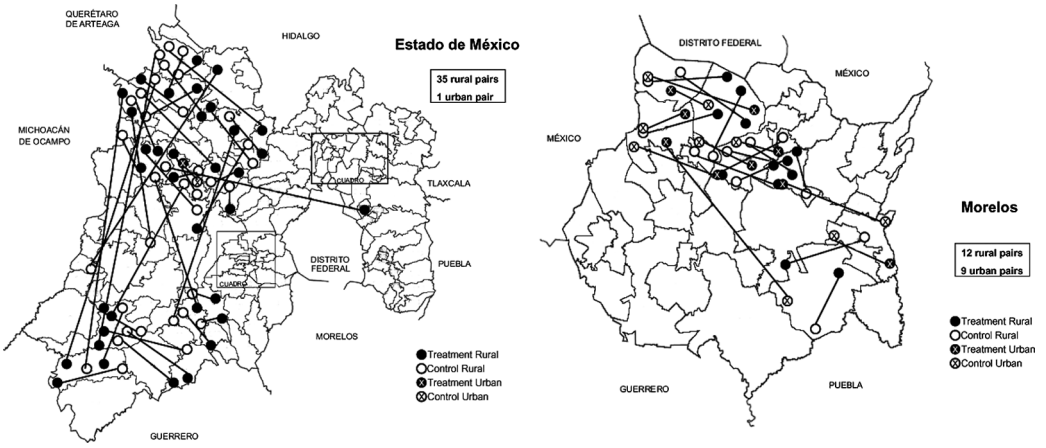

Figure 2. Matched pairs in Estado de México (left graph) and Morelos (right graph). Within states and an urban/rural distinction, health clusters were matched in pairs based on proximity to each other on various measured background characteristics. Graphs like this identify the exact clusters paired, and enabled us to use qualitative knowledge of these areas to check our quantitative matching algorithm.

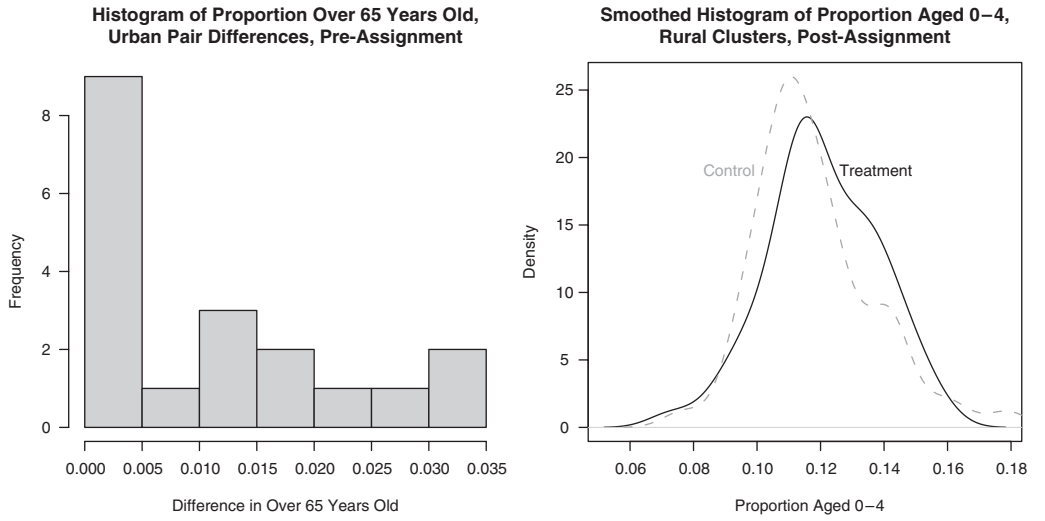

Figure 3. Balance in urban matched pairs in the over-65-year-old population, pre-randomization (left graph) and in the rural infant population, post-randomization (right graph). These graphs provide examples of the many checks we did to verify that our treated and control groups were similar on many measured background characteristics.

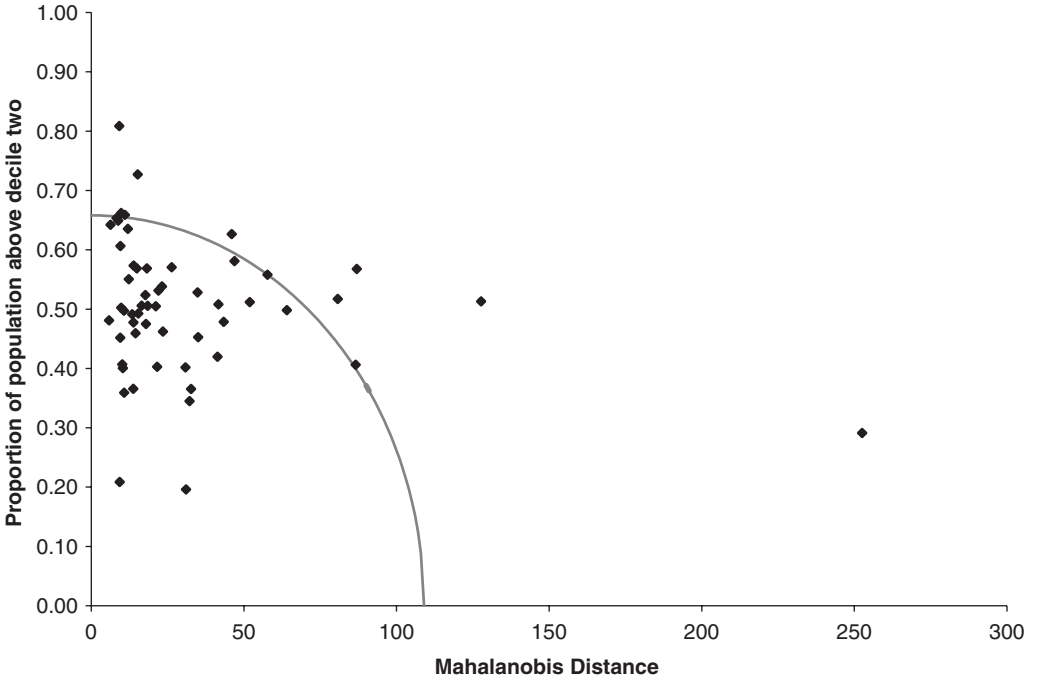

Figure 4. Choosing a subset of pairs for survey. We conducted our survey in 100 of the 148 health clusters with small Mahalanobis distances (pairs of clusters that were most similar) and low percentages above the first two deciles of income (where we expected highest compliance with our experiment). This graph portrays all rural pairs of clusters, for which we chose those represented as the points under the curved line.

reveals one badly matched pair all the way at the right, a few that are moderately bad in the middle, and that most are quite well matched on the left.

We also use the Mahalanobis distance and this same graph for a separate purpose. That is, although we conducted a survey of hospital facilities in every cluster in our study, we could only afford an individual-level survey in 100 of the 148 clusters. We decided that to keep enough power, our primary effort in this cohort of the experiment would be to estimate effects in rural areas and so we retained 90 of the 100 clusters from rural areas. We kept the remaining 10 urban clusters primarily so that our survey teams and the state and federal administrations would learn from the experience and be better prepared for future experimental cohorts of clusters in our ongoing evaluation. (We plan for the second cohort of our experiment to be from urban clusters.) To choose the particular 100 of 148 clusters, we used two criteria: the closeness of the match, measured via the Mahalanobis distance, and the probability of compliance with our randomized experimental encouragement, which we measured with the percentage of residents in the first two deciles of income (estimated from the 2002 National Income and Expenditure Survey). Figure 4 gives an example of the analysis we did for the rural clusters, with the Mahalanobis distance between elements of the pair on the horizontal axis and the percent above decile two in each pair on the vertical axis. Each dot then represents a pair of health clusters, and those in the lower left corner marked off by the curved line were pairs we chose to survey.

Random Assignment

After pairing clusters, we flipped the digital equivalent of a fair coin to choose one of the two clusters for treatment. Treatment was intended to include bringing the health clinic or hospital above a specified threshold level of doctors, specialists, nurses, equipment, office hours, technology, and drugs. It also was supposed to include setting up an MAO (Módulo de Atención y Orientación, or "service and orientation stand") in the health cluster, where citizens can go to affiliate, and advertising to encourage individuals to affiliate via radio, television, loudspeakers from cars, knocking on doors, painting walls (the Mexican version of billboards), or by other means. In addition, Mexican families enrolled in the Oportunidades antipoverty program, which comprises most of those in the lower two deciles of income, are affiliated automatically by the state. States receive funds only after affiliation is confirmed for each family, so motivation to encourage affiliation was strong. We also conveyed that we were more likely to be able to detect a positive effect of SPS, assuming one existed, if they did their best to affiliate in treatment clusters and to leave control clusters alone.

Encouragement efforts in our treatment clusters began in late August, 2005. Our survey began shortly thereafter. We plan to monitor affiliation efforts via studying the official Padrón, which is the confidential roster listing all persons affiliated and the trimesters they affiliated.

We also ran checks for the quality of both the matching and random assignment by examining overlapping histograms of treated compared to control clusters. The right graph in Figure 3 gives an example of a histogram (in the form of a kernel density estimate) for the proportion of population aged 0–4 years. The unbroken line gives the histogram for the treated group, and dashed is for the control group. As can be seen, the two are not identical, but they are close. These histograms are not identical because of the finite sample size and nonexact matches: As the number of health clusters with the same quality of matches grows, randomization guarantees that these histograms get closer and closer. Similarly, if clusters were available to produce exact matches, our matching algorithm would generate pairs of clusters that made these histograms the same.

Parametric Adjustment

Once the data are in, we need to compute a causal effect for every outcome variable. If matching is successful at balancing all potential confounding covariates, then a simple difference in means for an outcome measure between the treated and control groups would give an unbiased estimate of the causal effect of the policy decision to implement SPS at the level of the health cluster. Even if randomization fails, a difference in means could still give an unbiased estimate if the two groups happened to remain balanced on the observed background characteristics and any remaining imbalance were unrelated to the outcome variables. Similarly, if the randomization worked as designed, but we failed to measure and match on one or more important confounders or variables correlated with them, then the difference in means would still be unbiased. (And in either case, as described previously, we are protected from selection bias if we lose a cluster, to the extent that we matched on variables related to the reason for the loss.)

However, if *both* the randomization fails in some way and the matching was inadequate, then a simple difference in means between the control and treated groups can produce a biased estimate of the causal effect on the outcome variables measured in our surveys. Thus, if anything goes wrong and cannot be fixed with both of

these first two steps, we would drop the difference in means analysis. Instead, we would follow and adjust parametrically for any observed differences that may remain between the treated and control groups (Ho et al., 2007; Raudenbush, Martinez, & Spybrook, 2007). Thus, for outcome variables that are roughly continuous when aggregated to or measured at the level of the health cluster, a difference in means is equivalent to a linear regression of the outcome variable on the treatment indicator, with the coefficient on this indicator revealing the difference in means. To adjust parametrically, we would add to the regression any relevant pretreatment covariates, or functions thereof, that may still be confounders, possibly including interactions. If the parametric form is correct, bias will be reduced and the standard error will normally drop too. Other types of outcome variables would be analyzed by the relevant standard estimation approach and can include models for binary variables like logit, for event counts such as negative binomial regression, etc. (This procedure can even be made resistant to errors in the data introduced by political interventions we do not become aware of, or other problems, by using robust estimation techniques; Western, 1995; Zaman, Rousseeuw, & Orhan, 2001.) The right graph of Figure 3 gives an example of some small differences that remain in one of our background variables after matching and randomization that we adjust for parametrically.

As a result of this procedure, if either or both matching and randomization fail in some way, but the parametric specification adjusts appropriately for the relevant confounding variables or their correlates, then we can still obtain accurate estimates of the causal effects. Of course, this last step is a fail-safe, last resort technique, as fixing data problems by collecting better data is generally preferred to fixing them with assumption-based statistics after the fact (Wilde & Hollister, 2007). And valid randomization is still the only technique known to be able to avoid confounding from variables not measured or related to those matched on or adjusted for. Nevertheless, when planning experiments in a political environment, it pays to have this final piece of our triple robustness strategy available, because at least when the model is correct, appropriate bias-reducing adjustments can be made.

Design Limitations

Our evaluation design has several limitations that our subsequent analyses will have to deal with, in some cases via more sophisticated statistical procedures and in others via auxiliary data collection.

Most importantly, our clusters do not represent a random sample from the population of all clusters nationwide, and so generalization will need to await the results from new cohorts of our experiment. Of the 5,439 rural and 1,639 urban health clusters defined for the 13 states convinced to participate in the evaluation, we were able to retain 148 clusters in the study, including 55 rural and 19 urban pairs. Some pairs from each of 7 states are included, including Guerrero (1 rural and 6 urban), Jalisco (1 urban), Estado de México (35 rural, 1 urban), Morelos (12 rural, 9 urban), Oaxaca (3 rural, 1 urban), San Luis Potosí (2 rural), and Sonora (2 rural, 1 urban).

Figure 1 shows that the states in our experiment (in gray) are spread throughout the country. This diversity is useful both for generating a sample that is somewhat more representative and, especially for the states with fewer pairs, for helping us establish connections, communications, and practice with officials in these states for future cohorts of health clusters we intend to begin at later dates. However, many factors influenced their selection, only some of which we were able to observe and record. If other features of our evaluation design work as planned, we will have

unbiased estimates for these areas, but further research, survey comparisons with national statistics, and subsequent waves of our experiment are required before we can ascertain whether results we find apply more broadly. We thus followed the medical model of maximizing the chances that our random assignment would be executed as planned, so that inferences for the sample at hand are valid, even though selection into the sample was not randomized or fully controlled.

We can briefly compare our sample with that from ENSANut 2005 (Encuesta Nacional de Salud y Nutrición, a national survey of 45,241 Mexican adults, to give a sense of the areas in our evaluation. The single biggest difference between the two surveys is that our baseline survey has an (intentional) rural bias, given that 90 out of the 100 clusters we chose to include in the study are rural, whereas the nationally representative sample of the ENSANut is approximately the opposite. Only 10 percent of the households in the baseline are from urban areas, whereas 77 percent of the households in the ENSANut are drawn from urban areas. The demographic compositions of the two samples are otherwise fairly similar. The ratio of male to female heads of household is almost the same in the two samples, as are the distribution of education, and age composition of the primary respondents.

Other design limitations include the fact that we were unable in our prerandomization matching process to control for the proximity of our control clusters to treatment clusters, or other clusters in which SPS was already in operation, and so we will need to check for any spillover effects and correct for them if necessary. More detailed verification will be useful, such as verifying from the *Padrón* how many citizens affiliated in each of our health clusters and how much use they made of SPS medical services. The level of encouragement used in different clusters may also have varied in ways we were unable to monitor.

Although our design is protected from selection bias when losing clusters due to reasons related to the variables used to create the matched pairs, we only have 148 clusters in total (and 100 in which our individual-level survey was conducted), and so we risk having little power if we lose too many. Our experiment contains many outcomes, which is valuable, but also risks a "multiple comparisons" problem if not analyzed properly; publicly stated ex ante theoretical expectations will ameliorate some of this problem, but, as the next section details, disagreement about likely specific outcomes requires that some of this problem will need to be addressed by statistical procedures during the analysis stage. And finally, although our evaluation design is robust to some types and degrees of political intervention, no design can avoid all such problems. Indeed, like most evaluations, ours could be terminated at any time by the same government officials now facilitating its continuation or by the next elected government.

EXPECTED PROGRAM EFFECTS

In order to ascertain the *intended* goals of SPS, we convened, on three separate occasions, large meetings of political appointees, administrators, and local experts from the federal and state governments. We elicited from these individuals and groups, in a variety of qualitative and quantitative ways, where and when they thought SPS would be likely to succeed and fail. Initially, we attempted to pin down individual quantitative predictions by giving them lists of the outcome variables with likely confidence intervals sizes for the evaluation. If this worked, we would then "tie our own hands" and explain at that time, before the data came in, exactly what analyses we would run when these data eventually became available and what we would conclude if the results turned up in different ways.

Although these meetings were informative, our strategy did not work as planned. The "Mexican government" is no more a unitary actor with a single opinion than any other government, and the groups that marshaled support for, passed, run, and are responsible for SPS are far too diverse to expect them to give precise or even qualitatively similar answers to our quantitative questions. We thus abandoned this strategy and instead report here our qualitative understanding of the government's expectations.

The groups we spoke with were nearly unanimous on the importance of the program and its likelihood of eventual and considerable overall success, but there was disagreement about how the program would have its effects, which effects would be detectable given the likely sizes of our confidence intervals, how long the program would take to start showing health effects on the population, and in which regions or areas the program would have the biggest effect. Other disagreements were effectively based on different theories of individual behavior, how much of the funds would get to the clinics and people who need it, and the likely compliance behavior of the Mexican population with medical advice.

For one example, because a central feature of the program is financial protection from catastrophic health expenditures, many believed this would work, but some thought that it would not be easily detectable in the short run. The source of this disagreement was based on different understandings of how huge medical expenses now affect the population prior to SPS. Some believe that citizens who are suddenly hit with some very expensive medical payment have a similarly sudden and large reduction in their nonhealth disposable income. Others believe that people instead find partial solutions that they can afford. So, for example, when having a child, instead of "selling the farm" to pay for a stay at the hospital miles away and all the associated care with OB/GYN physicians and specialized equipment, the idea is that people without much income instead opt for a less expensive midwife and so do not incur a catastrophic expenditure. If the latter is true, then SPS will improve care and reduce family expenditures, but the effect will not be as large or dramatic and so may not be as easy to detect. And still others are mainly focused on catastrophic expenditures that come from expensive medicines paid for over a longer period of time.

If SPS is to be a success, the initial unambiguous sign will be that utilization of medical services will increase. The number of visits to health clinics, doctor visits, medicines prescribed, etc., should increase, as should the number of medical diagnoses made. Individual health expenditures should drop, including total out-of-pocket spending, catastrophic expenditures (paying more than 30 percent of disposable income on health), and impoverishment due to health care payments (households pushed below the poverty line because of health care spending). If SPS is an effective program, we would expect to see these changes relatively quickly, although perhaps not all by 10 months. We expect most other causal effects of SPS estimated not to show detectable effects by a mere 10 months, but we decided to measure many others (summarized in a section above) in order to collect baseline information, as a check on our design, and to establish a framework to monitor conditions in the long run. The effects of SPS on a few of these other measures might also conceivably be detectable in our first follow-up survey.

EMPIRICAL VALIDATION

As a supplement to our triply robust evaluation design, and our paired matching that protects us in some circumstances from selection bias even if we lose some health clusters, we now report an empirical check of the validity of all the steps

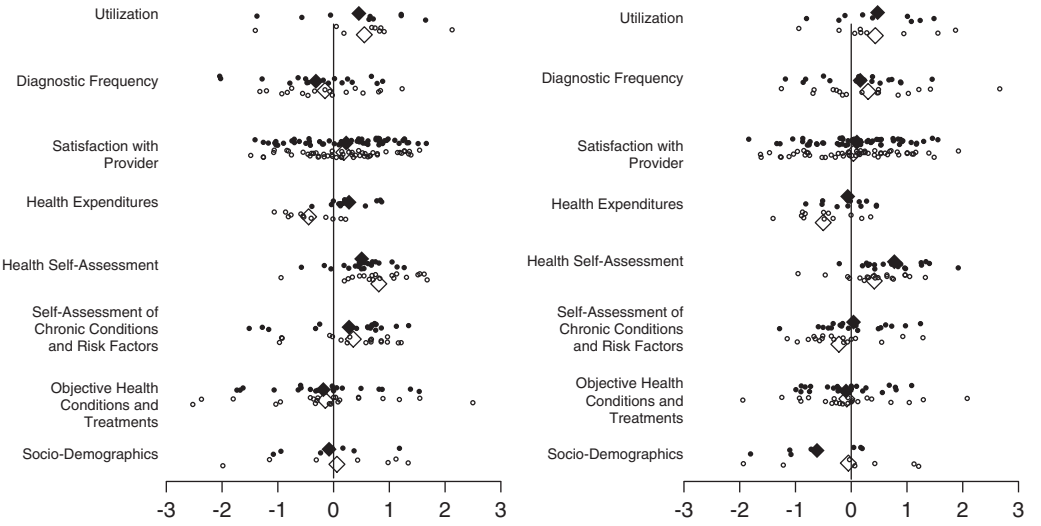

Figure 5. Effects of random assignment on outcome measures at baseline, for all families (left graph) and poor families, in Oportunidades (right graph). If the experiment were implemented properly, we would see zero effect (near the vertical line) plus or minus random error. The horizontal axis is in standard deviation units, and so we expect relatively few estimates outside the $[-2, 2]$ interval, for example, which appears to be the case. Estimates appear without (in open circles) and with (closed disks) covariate adjustment; corresponding diamonds represent the average for each category.

(taken together) that we have implemented thus far. To do this, we estimate the causal effect of treatment assignment on a large number of our outcome variables measured in our baseline survey, standardized to the same scale (by dividing by the standard error).[10]

We present these analyses, for all of our 31,856 respondents, in the graph at the left of Figure 5. This graph gives causal effect estimates for our variables organized into eight categories (with individual items listed in the first appendix.[11] Each effect is estimated twice, once without parametric correction (on the graph in open circles) and once with it (the black disks). The corresponding open and closed diamonds are the average for each category. The horizontal axis is denominated in standard deviation units. If all the outcome measures were independent, we would expect 95 percent of the points on the graph to be between -2 and 2. The outcome measures are surely not independent, but most are indeed in this interval and, with one partial exception (discussed below), all the averages within categories are fairly close to zero.

Similar results appear in the right graph, which uses the same analytical procedures applied to low-income respondents (in Oportunidades), and in the two graphs in Figure 6 for relatively more wealthy families (those with formal sector health insurance and/or a large asset count) on the left, and those who are neither poor nor

[10] We analyzed these data by multiply imputing the relatively small fraction of missing data at the individual level, aggregating each variable to the cluster level, analyzing the completed data sets as described above, following standard procedures for combining the separate analyses from each imputed data set (see Honaker and King, 2006; King et al., 2001; Rubin, 1987), and translating the coefficients from the various models into the quantities of interest (King, Tomz, & Wittenberg, 2000; Imai, King, & Lau, 2006).
[11] All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

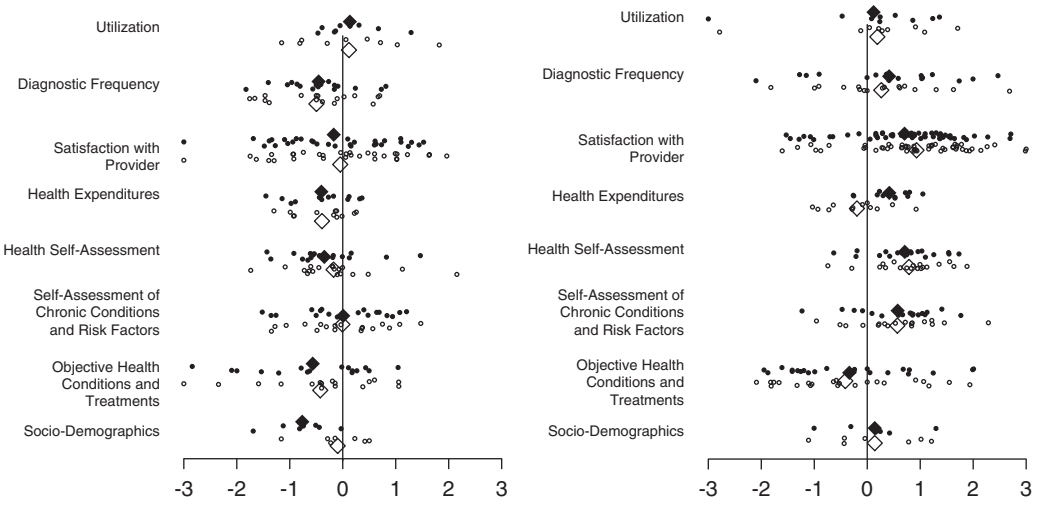

Figure 6. Effects of random assignment on outcome measures at baseline, for relatively wealthy families (left graph) and those who are neither poor nor wealthy (right graph). See the caption to Figure 5 for details.

wealthy (not in Oportunidades, no formal sector insurance, and no high asset count) are on the right.

The one exception to complete confirmation of the success of our design is in the category of health self-assessments. Although almost all the dots even in this category are within the −2 to 2 interval, there appears to be a slight pattern with treatment groups apparently causing the poor to report that they are healthier and the more wealthy to report that they are sicker. Yet, individuals in the treated group have not received any treatment other than random assignment to be encouraged to affiliate and knowledge of this assignment. If we find in the follow-up survey an effect that is no larger than the one here, we know now to attribute it to measurement problems, such as "differential item functioning" (Holland & Wainer, 1993; King, Murray, Salomon, & Tandon, 2004) rather than to a true causal effect.[12] This result also reveals an important benefit of fielding the baseline survey: If we had no baseline survey, and this pattern had appeared in the follow-up, we might have incorrectly concluded that SPS was making the poor healthier but the rich, who were enrolled in other insurance plans, less healthy.

[12] Because the estimates are correlated, it may be that this pattern is a random occurrence. If not, it might be a Hawthorne effect, because many of the respondents in the group who would eventually be encouraged to affiliate with SPS were aware of this at the time of the survey. Lower income citizens, who would benefit from the program, would by this account be conveying with their "biased" responses that they would not be a burden on the system if they were given access to SPS. In contrast, those with more income, who would likely keep their existing health insurance even if in the treated group, might not favor the government spending a lot of money on a program they would not benefit from, and so they may be communicating to some degree that they are in need of more help than the government is planning to provide them. Officials indicate that this type of pattern has occurred before in response to government programs in Mexico.

We studied the analyses for the few other dots outside the −2 to 2 interval in these graphs for other categories and did not find any systematic or patterns that seem troublesome. They appear to be random occurrences, which we would expect for some fraction of the estimates, even if the true effect were exactly zero.

CONCLUDING REMARKS

This evaluation is a rare opportunity to learn about and improve a public policy program in which Mexico is investing a great deal of time, money, and effort. A key to the evaluation is that it is being conducted without delaying the implementation of the program or slowing what the government views as its attempt to give millions of people healthier and longer lives, free from health spending-induced financial impoverishment. We do not know how Seguro Popular, or its many components, will be evaluated in the end, but we are certain that thousands of national and regional governments around the world, as well as their citizens, would greatly benefit by following the lead of the Mexican government and enabling social scientists to conduct serious, arms-length, dispassionate, scientific evaluations of governmental programs.

In return, as scientists, we must understand, accommodate, and adapt to the political realities in which governments and policymakers operate. High-minded science that is not designed to fit in local politics risks accomplishing little of practical value. In addition to reporting on how we conducted this evaluation, we have attempted in this paper to offer some methods that may make it possible for others to design politically robust evaluations of a diverse array of different public policy programs. We hope future researchers will be able to build on these techniques and develop others so that policy experiments eventually become almost as common as new public policy programs.

We believe that aspects of our "politically robust" experimental design should be widely applicable in other policy evaluation settings, particularly in the developing world. We know this should be possible because we adapted most parts of our design from components that have already been used in previous evaluations. Cost should also not be a concern in future evaluations: our project is unusually large compared to previous efforts, but the total cost, the bulk of which is due to the expense of running large surveys, is a tiny fraction of the cost of the program itself. If we are able to improve future administration of SPS in only minor ways, learn that SPS should continue to be rolled out in the same way as it has been already, or find that the program has failed and so funds can be redirected faster, the return on investment in terms of the financial and health benefits to the citizens of Mexico should be orders of magnitude larger than the cost of the evaluation.

The main intended contribution of this paper, in addition to a variety of specific technical suggestions, is the perspective of designing field experiments that are capable of surviving the problems that we can all expect will naturally occur in the real world. In addition to the problems generated routinely in democratic systems that we have focused on, it would also be worthwhile for future researchers to consider how to produce evaluation methods that can survive many other types of problems as well, such as due to logistical, administrative, technical, and implementation issues; cultural mishaps; natural and other disasters; and the whole range of compliance problems. We hope future researchers will work on continuing to develop new fail-safe evaluation methods, so that the remarkable power of experimental designs can be fully brought to bear on the problems that affect human populations.

GARY KING is the David Florence Professor of Government, and the Director of the Institute for Quantitative Social Science, Harvard University.

EMMANUELA GAKIDOU is a Research Associate at the Institute for Quantitative Social Science, Harvard University.

NIRMALA RAVISHANKAR is a graduate student affiliate at the Institute for Quantitative Social Science, Harvard University.

RYAN T. MOORE is a graduate student affiliate at the Institute for Quantitative Social Science, Harvard University.

JASON LAKIN is a graduate student affiliate at the Institute for Quantitative Social Science, Harvard University.

MANETT VARGAS was Research Manager of the Mexican Health System Evaluation Project at the Institute for Quantitative Social Science, Harvard University, during this project and is now Acting General Director of the Oportunidades Program, National Commission for Social Protection in Health, Ministry of Health, Mexico.

MARTHA MARÍA TÉLLEZ-ROJO is Director of Human Ecology, Instituto Nacional de Salud Pública (National Institute of Public Health), Mexico.

JUAN EUGENIO HERNÁNDEZ ÁVILA is Director of Information and Medical Geography, Instituto Nacional de Salud Pública (National Institute of Public Health), Mexico.

MAURICIO HERNÁNDEZ ÁVILA was General Director of the Instituto Nacional de Salud Pública (National Institute of Public Health), Mexico, during this project and is now Undersecretary for Prevention and Health Promotion, Secretaría de Salud (Ministry of Health), Mexico.

HÉCTOR HERNÁNDEZ LLAMA was Coordinator of the Supply of Health Services, Secretaría de Salud (Ministry of Health), Mexico, during this project and is now a consultant at Conestadistica.

ACKNOWLEDGMENTS

Our thanks to Octavio Gómez Dantés and Sergio Sesma for much helpful advice throughout the project; René Santos Luna for help in constructing health clusters; Manuel Castro for managing the INSP survey team; Ferdinand Alimadhi and Elena Villalon at IQSS for statistical programming; Eduardo Lazcano for support and information; Jesse Abbott-Klafter, Chunling Lu, Chris Murray, Emre Ozaltin, and Cecilia Vidal for suggestions on the questionnaire; Jeremy Barofsky, Chloe Bryson-Cahn, Dennis Feehan, and Diana Lee from Harvard and Maritza Solano Gonzalez, Aaron Salinas Rodriguez, and Francisco Javier Carlos Rivera from INSP for research assistance; our formally appointed panel of experts, Edmundo Berumen, Luis Felipe Lopez Calva, Nora Claudia Lustig, Thomas Mroz, and John Roberto Scott for many helpful suggestions; Jim Alt, Mitchell Duneier, Don Green, Kosuke Imai, Steve Kelman, Joe Newhouse, and Ken Shepsle for helpful advice; Howard Bloom, a (formerly anonymous) reviewer, for his generous help and insight; and the National Institute of Public Health of Mexico, the Mexican Ministry of Health, the National Institutes of Aging (P01 AG17625-01), and the National Science Foundation (SES-0318275, IIS-9874747, SES-0550873) for research support.

REFERENCES

Adato, M., Coady, D., & Ruel, M. (2000). An operations evaluation of PROGRESA from the perspective of beneficiaries, promotoras, school directors and health staff. Final report, International Food Policy Research Institute, 2033 K Street, NW Washington, DC 20006.

Alesina, A., & Tabellini, G. (1990). A positive theory of fiscal deficits and government debt. The Review of Economic Studies, 57, 403–414.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. Journal of the American Statistical Association, 98, 299–324.

Behrman, J. R., & Todd, P. E. (1999). Randomness in the experimental samples of PROGRESA (education, health, and nutrition program). Research report. Washington, DC: International Food Policy Research Institute.

Bloom, H. S., Richburg-Hayes, L., & Black, A. (2007, in press). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis.

Bohm, P. (1984). Are there practicable demand-revealing mechanisms? In H. Hanusch (Ed.), Public finance and the quest for efficiency (pp. 127–139). Detroit: Wayne State University Press.

Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., de Moya, D., Grimshaw, J., & Foley, E. (2004). Estimating the effects of interventions that are deployed in many places: Place-randomized trials. American Behavioral Scientist, 47, 608–633.

Boruch, R. F. (1997). Randomized experiments for planning and evaluation. Thousand Oaks, CA: Sage Publications.

Box, G. E., Hunger, W. G., & Hunter, J. S. (1978). Statistics for experimenters. New York: Wiley-Interscience.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. The Journal of Economic Perspectives, 9, 63–84.

Camasso, M. J., Jagannathan, R., Harvey, C., & Killingsworth, M. (2003). The use of client surveys to gauge the threat of contamination in welfare reform experiments. Journal of Policy Analysis and Management, 22, 207–223.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally College Publishing Company.

Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. Journal of Policy Analysis and Management, 23, 471–488.

Derthick, M. (1979). Policymaking for Social Security. Washington, DC: The Brookings Institution.

Donner, A., & Klar, N. (2000). Design and analysis of cluster randomization trials in health research. London: Arnold.

Flay, B. R., & Best, J. A. (1982). Overcoming design problems in evaluating health behavior programs. Evaluation & The Health Professions, 5, 43–69.

Frangakis, C. E., Rubin, D. B., & Zhou, Z.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. Biostatistics, 3, 147–164.

Franzese, R. (2002). Macroeconomic policies of developed democracies. New York: Cambridge University Press.

Frenk, J., Gonzalez-Pier, E. Gomez-Dantes, O., Lezana, M. A., & Knaul, F. M. (2006). Comprehensive reform to improve health system performance in Mexico. The Lancet, 368, 1524–1534.

Frenk, J., Sepúlveda, J., Gómez-Dantés, O., & Knaul, F. (2003). Evidence-based health policy: Three generations of reform in Mexico. The Lancet, 362, 1667–1671.

Gertler, P. (2006). Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. The American Economic Review: Papers and Proceedings, 94, 336–42.

Gertler, P. J. (2000). Final report: The impact of PROGRESA on health. International Food Policy Research Institute.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. Journal of Policy Analysis and Management, 25, 75–96.

Goodman, J. S., & Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. Journal of Management, 22, 627–652.

Green, D. P., & Gerber, A. S. (2002). Reclaiming the experimental tradition in political science. In Milner, H., & Katznelson, I. (Eds.), State of the discipline, vol. III (pp. 805–832). New York: W.W. Norton & Company, Inc..

Greenberg, D., & Shroder, M. (2004). The digest of social experiments (3rd ed.). Washington, DC: Urban Institute Press.

Greenberg, D. H., Michalopoulos, C., & Robins, P. K. (2006). Do experimental and nonexperimental evaluations give different answers about the effectiveness of government funded training programs? Journal of Policy Analysis and Management, 25, 523–552.

Greevy, R., Lu, B., Silver, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. Biostatistics, 5, 263–275.

Harrison, G. W., & List, J. A. (2004). Field experiments. Journal of Economic Literature, 42, 1009–1055.

Heckman, J. J. (1992). Randomization and social policy evaluation. In Manski, C. F., & Garfinkel, I. (eds.), Evaluating welfare and training programs. Boston: Harvard University Press.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. The Journal of Economic Perspectives, 9, 85–110.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics, 1, 69–88.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for parametric causal inference. Political analysis. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Holland, P. W., & Wainer, H. (Eds.) (1993). Differential item functioning. Hillside, NJ: Lawrence Erlbaum.

Holzer, H. J., Quigley, J. M., & Raphael, S. (2003). Public transit and the spatial distribution of minority employment: Evidence from a natural experiment. Journal of Policy Analysis and Management, 22, 415–441.

Honaker, J., & King, G. (2006). What to do about missing values in time series cross-section data. http://gking.harvard.edu/files/abs/pr-abs.shtml.

Howell, W. G. (2004). Dynamic selection effects in means-tested, urban school voucher programs. Journal of Policy Analysis and Management, 23, 225–250.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. American Political Science Review, 99, 283–300.

Imai, K., King, G., & Lau, O. (2006). Zelig: Everyone's statistical software. http://gking.harvard.edu/zelig.

Imai, K., King, G., & Stuart, E. (2007). Misunderstandings among experimentalists and observationalists: Balance test fallacies in causal inference. http://gking.harvard.edu/files/abs/matchse-abs.shtml.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review, 95, 49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review, 98, 191–205. http://gking.harvard.edu/files/abs/vign-abs.shtml.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. American Journal of Political Science, 44, 341–355. http://gking.harvard.edu/files/abs/making-abs.shtml.

Kish, L. (1949). A procedure for objective respondent selection within the household. Journal of the American Statistical Association, 44, 380–387.

Klar, N., & Donner, A. (1997). The merits of matching in community intervention trials: A cautionary tale. Statistics in Medicine, 16, 1753–1764.

Klarman, M. J. (1997). Majoritarian judicial review: The entrenchment problem. The Georgetown Law Journal, 85, 491–554.

Kramer, M., & Shapiro, S. (1984). Scientific challenges in the application of randomized trials. Journal of the American Medical Association, 252, 2739–45.

Krueger, A. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 114, 497–532.

Lakin, J. (2005). Letting the outsiders in: Democratization and health reform in Mexico. Paper presented at the annual meeting of the American Political Science Association, Washington DC.

Lewis, M. (2005). Improving efficiency and impact in health care services: Lessons from Central America. In Forgia, G. M. L. (ed.), Health systems innovation in Central America. Washington, DC: The World Bank.

Murray, C. J., & Evans, D. B. (Eds.) (2003). Health systems performance assessment: Debates, methods and empiricism. Geneva: World Health Organization.

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

Nickerson, D. W. (2005). Scalable protocols offer efficient design for field experiments. Political Analysis, 13, 233–252.

Palmer, T., & Petrosino, A. (2003). The "experimenting agency." The California Youth Authority Research Division. Evaluation Review, 22, 228–266.

Posner, E. A., & Vermeule, A. (2002). Legislative entrenchment: A reappraisal. The Yale Law Journal, 111, 1665–1705.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. Educational Evaluation and Policy Analysis.

Robins, J. M., & Rotnitzky, A. (2001). Comment on the Peter J. Bickel and Jaim Young Kwon, "Inference for semiparametric models: Some questions and an answer." Statistica Sinica, 11, 920–936.

Rosner, B., & Hennekens, C. H. (1978). Analytic methods in matched pair epidemiological studies. International Journal of Epidemiology, 7, 367–372.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley.

Sommer, A., & Zeger, S. (1991). On estimating efficacy from clinical trials. Statistics in Medicine, 10, 45–52.

Sterk, S. E. (2003). Retrenchment on entrenchment. The George Washington Law Review, 71, 231–254.

Thompson, D. F. (2005). Democracy in time: Popular sovereignty and temporal representation. Constellations, 12, (pp. 245–261).

Torp, H., Rauum, O., Hernaes, E., & Goldstein, H. (1993). The first Norwegian experiment. In Karsten, J., & Madsen, P. K. (eds.), Measuring labour market measures: Evaluating the effects of active labour market policies. Copenhagen, Ministry of Labour.

Western, B. (1995). Concepts and suggestions for robust regression analysis. American Journal of Political Science, 39, 786–817.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class-size reduction experiment. Journal of Policy Analysis and Management, 26, 455–480.

Zaman, A., Rousseeuw, P. J., & Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. Economics Letters, 71, 1–8.

Appendix A. Outcome Measures.

This appendix lists the dependent variables in our surveys used for estimating causal effects in Figures 5 and 6 in their corresponding categories, followed by other variables we also collected. All data, codebooks, and replication information from this project that we are legally permitted to distribute (that is, excluding items like the Padrón) will be made publicly available upon publication.

Utilization: Health insurance, SPS affiliation, health care available when needed, number of prescribed medicines able to get, ease in getting needed medications, days/week and hours/day health clinic is open, inpatient and outpatient visits.

Diagnostic Frequency: Diagnosed, treated, and presently taking medicines for arthritis, heart disease, asthma, depression, diabetes; hypertension and hypercholesterolemia diagnoses; vision difficulties.

Satisfaction with Provider Difficulties with health care providers, quality of SPS services, satisfaction with quality SPS services, selection of and quality of services from Instituto de Mexicano del Seguro Social (IMSS), IMSS-Oportunidades, Instituto de Sequridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE), SPS, PEMEX. For both inpatients and outpatients: traveling and waiting time; cleanliness of facilities; whether talked to respectfully and privately; received clear explanations; had enough time for questions; involved in decisions; confidentiality; freedom to choose provider; adequate space; treated worse by health care provider because of sex, age, lack of money, social class, ethnic group, illness, other.

Health Expenditure: annual out of pocket health expenditures in many categories; measures of catastrophic health expenditures (for example, greater than 30 percent of disposable income).

Health Self-Assessment: Self-assessment of overall health; difficulty moving around; in vigorous activities; self-care; maintaining general appearance; feeding self; bodily aches or pain, soreness or discomfort; problems in daily life due to pain, concentrating or remembering things, learning a new task, personal relationships or participation in community; getting along with others; performing work or other regular daily activities; sleeping; not feeling rested and refreshed during the day; feeling sad, low, or depressed; problem with worry or anxiety; general satisfaction with health.

Self-Assessment of Chronic Conditions and Risk Factors: Smoking at all or daily; drinking alcohol and amount; eating fruits; eating vegetables; joint pain, aching, stiffness or swelling; stiffness in joint in morning or after long rest, joint pain goes away after exercising or movement; back pain; discomfort in chest when walking uphill; discomfort in chest when walking; attacks of wheezing; tightness in chest; shortness of breath without obvious cause when not engaging in physical activity; depression.

Objective Health Conditions and Treatments: Coverage for antenatal care; acute respiratory infections for children; systolic blood pressure, hypertension control, diagnosis, and treatment; cervical exam; cholesterol level, control, diagnosis, and treatment; diarrhea for children coverage; diabetes control coverage, diagnosis, treatment; flu vaccine; glasses; high cholesterol; hypertensive; mammography; seeing health care professional during pregnancy; Pap smear coverage; skilled birth attendance coverage.

Sociodemographics: weight, height, marital status, education, attend religious services, employment, reason for unemployment.

Other Variables: Dwelling characteristics (material of floors, ceiling, walls, number of rooms), access to services (electricity, sewage, etc), and assets owned by household; satisfaction with SPS affiliation process; social capital and stress (feeling of security, violence, opinion on main problem faced by the country, opinion on who is

responsible for problems in the health sector, opinion on who is responsible for the creation of SPS, opinion on who should pay for health services; frequency of access to news on TV, radio, newspapers, trust in media, ideological position, opinion on Mexico's economic, political and social situation).

Appendix B. Analysis Plans.

When each wave of post-treatment data come in, we plan to conduct analyses at two levels, each involving more sophisticated statistical analyses. As Imai (2005) writes, "If field experiments work perfectly … and empirical relationships are unambiguously strong, then sophisticated statistical analysis may be unnecessary. However, precisely because field experiments take place in the real world, such perfection is almost never achieved in practice."

Our first analyses from both the household and facilities surveys will be at the level of the health cluster and will be conducted in a manner analogous to that in the Empirical Validation section above. For variables aggregated up from the individual survey data to the cluster level, we will multiply impute item nonresponse as well as some entire survey responses due to the expected 8–10 percent attrition rate for Mexican surveys like these (which is relatively low compared to surveys in the U.S.; for example, Holzer, Quigley, & Raphael, 2003). In addition, the specific imputation techniques we use will need to take account of the fact that compliance with the experiment is estimable with appropriate models, but not predictable from standard imputation approaches (see Hirano et al., 2000).

We will also need to compensate for unit nonresponse and the resulting selection problems that may occur, such as those who are ill and do not feel well enough to participate, and those who have died, who will obviously not participate. Although sample attrition is usually ignored in experiments (Goodman & Blum, 1996), doing so can generate considerable bias (for example, Sommer & Zeger, 1991). Ignoring missing data, such as via listwise deletion, or imputing assuming standard "missing at random" assumptions would thus bias our evaluation, and so statistical techniques designed for these problems are necessary. In addition, we will search for evidence that SPS is working better in some areas than others, and try to characterize what it is about those areas that might breed success (for example, Rosner & Hennekens, 1978). The leading hypothesis going in is that SPS is more effective in poorer areas. Although we have only 50 matched pairs, we might be able to detect these differences by simply dichotomizing the same and running the same analyses.

The other key analysis to be conducted will be at the individual level, where we attempt to estimate the individual level causal effect affiliation with SPS, and the associated medical and financial services made available, on the health and well-being of individuals who comply with the encouragement assignment. Compliance issues are the key statistical problem here, as we could not randomize individuals to SPS affiliation. We instead randomized encouragement (and the funds for available health care), and so compliance with our encouragement must be estimated. It turns out to be possible to estimate the effect of SPS using our design on compliers (that is, those who affiliate because they are encouraged in our treatment groups and who do not affiliate because of the lack of encouragement in our control groups), and for other groups of interest (Hirano et al., 2000; Barnard et al., 2003). We also have an advantage over other applications of the same ideas, because all those who were enrolled in Oportunidades will be affiliated to SPS automatically, although we will have to ascertain the extent to which these individuals are aware of their affiliation.

Our intent-to-treat causal estimates should be of interest to policymakers, especially in the states, deciding whether and how to roll out the program in new areas. The individual-level causal estimates should be of interest to both policymakers and public health officials as they try to improve the operation of the program and, ultimately, the health of the people. Throughout, we hope to find clues about what works, what does not work, and most importantly, ways of improving the structure, organization, operation, and focus of the SPS program.