Methodological Frontiers of Public Finance Field Experiments

Jeffrey R. Kling*

February 2007

National Tax Journal, forthcoming

* The Brookings Institution and NBER.

I thank Rosanne Altshuler, Greg Duncan, William Gale, Lisa Gennetian, Lawrence Katz, Erzo Luttmer, and Sendhil Mullainathan for comments and suggestions. I gratefully acknowledge support for this work provided by the John D. and Catherine T. MacArthur Foundation, the Charles Stuart Mott Foundation, and the William T. Grant Foundation.

February 2007

METHODOLOGICAL FRONTIERS OF PUBLIC FINANCE FIELD EXPERIMENTS

Jeffrey R. Kling

ABSTRACT

The purpose of this article is to demonstrate how a rich array of methods can be applied to increase the relevance of field experiments in public economics. Two cross-cutting themes are important in multiple phases of the research. First, greater statistical sophistication can draw more value from a field experiment without obscuring the simple and compelling information from the differences in average outcomes of intervention and control groups. Second, the methodological frontier is interdisciplinary, drawing on knowledge and techniques developed in psychology, anthropology, and sociology that can be adapted in order to make public finance field experiments more useful.

Keywords: social experiment; program evaluation; mixed methods

JEL classifications: C93, H0, I0, J18

Jeffrey R. Kling
The Brookings Institution
1775 Massachusetts Avenue
Washington, DC 20036
jkling@brookings.edu

An economics graduate student starting her dissertation today would do well to follow the example of Heather Ross and think big. After working in 1965 and 1966 as an intern with the President's Council of Economic Advisers and the U.S. Department of Health, Education and Welfare, Ross became interested in income maintenance policy, particularly in how the behavior of low income people might respond to transfer payments. In 1966, Ross, then a visiting research fellow at the Brookings Institution, wrote a paper entitled *A Proposal for Demonstration of New Techniques in Income Maintenance,* which was subsequently submitted to the Johnson Administration's anti-poverty agency, the Office for Economic Opportunity. The experiment that resulted from this proposal, the New Jersey Income Maintenance experiment, is generally considered to be the first large-scale social experiment ever conducted. Ross's thesis proposal instigated a project that cost over $30 million in 2006 dollars, and sparked a wave of large-scale experiments in health, housing, and welfare.[1]

Over the past forty years, many of the essential goals of public finance economists in assessing policy interventions have remained the same as they were for Ross. We would like to have a credible strategy for identifying the causal effect of the intervention, and we would like to have data about situations that are as similar as possible to that in which the policy would occur in the future. In the case of income maintenance programs, there was little data in 1966 that could be used to meaningfully forecast the potential effects of the negative income tax proposals that were being made. The compelling logic of Ross's proposal was that in addition to collecting contextually relevant data, one could use the power of random assignment in an experimental framework to assess how otherwise similar groups of individuals responded when assigned to

[1] Ross received her PhD in economics from the Massachusetts Institute of Technology in 1970 and subsequently held various posts in government including Special Assistant to the President for Economic Policy. She is currently a visiting scholar with Resources For the Future. The historical account of her dissertation and its influence is based on Greenberg and Shroder (2004) and Orr (1999). The results of that and other early subsequent experiments were reviewed by Ferber and Hersh (1978).

different policy interventions. The difference in outcomes between an intervention group and a comparison group in a social experiment could be used as an estimate of the effect of the intervention in a setting similar to that in which a policy intervention might be enacted.

The purpose of this article is to demonstrate how a rich array of methods can be applied to increase the relevance of field experiments in public economics. Given that the essence of analyzing experiments typically involves looking at the difference between two sample means, one might ask whether experiments involve more than rudimentary research methods, and even whether there are any important methodological frontiers in this area. I focus on advances in the research methods used in conducting field experiments, using examples from government taxation and expenditure, public goods, and externalities. The examples are intended to illustrate methods with general applicability, and this is not an exhaustive survey of particular studies. In this article I assume that the reader is familiar with the basics of how to design an experiment and the conditions under which it identifies the causal effect of a policy intervention, as discussed by Orr (1999) and others. The focus here is on research in North America.[2] To emphasize frontiers of research, the examples are primarily drawn from studies using data collected within the past few years.

To characterize the public finance field experiments I discuss, the circumstances are generally intended to be close to an actual application of a policy intervention as possible – although some are clearly closer than others.[3] In focusing on public finance field experiments, I also adopt criteria delineated by Greenberg and Shroder (2004) for identifying social

[2] For discussion of field experiments in developing countries, including some relating to public finance, see Duflo (2006).

[3] Harrison and List (2004) contrasted lab and field experiments in terms of the subjects, the information they bring to the task, the nature of the commodities, the tasks, the stakes, and the environment for the subjects to develop a taxonomy of experiments. In their taxonomy, experiments with a non-student subjects and non-lab contexts for the commodity, task, or information set of subjects are referred to as framed field experiments. A framed field experiment where the subjects do not know they are in an experiment is referred to as a natural field experiment. In this taxonomy, I discuss include both framed and natural field experiments.

experiments, which are: random assignment, policy intervention, follow-up data collection, and analysis. I use the term 'field experiments' here mainly because some authors, including Ferber and Hirsch (1978), have defined 'social experiments' as those being conducted with government funds or by government agencies. For the purposes of this article, I will consider all social experiments to be public finance field experiments.

In this article I will not focus on the lengthy and informative debate over the value of public finance field experiments relative to other types of research. Many of the potential issues were identified early on by Rivlin (1974) and others. For instance, an experiment may not be a good estimate of the effects of a policy if people behave differently when they know they are being observed, if they believe the experimental intervention is temporary but the policy would be permanent, or if the policy created spillovers when implemented at scale that were not present in the experiment. Each experiment attempts to minimize the impacts of such issues, and naturally there is variation in the success of these attempts. The issues have been ably discussed at length by Burtless (1995), Heckman and Smith (1995), Harrison and List (2004), Moffitt (2004) and others. Rather than reviewing this terrain again, I will start from the premise that most experiments have value and focus on methodological approaches to making public finance field experiments more useful.[4]

This article is organized around the sequence of activities involved in conducting an experiment and discusses related methodological innovations. First, topics are selected. Second, a conceptual framework is developed that motivates hypotheses about the potential impacts of

[4] I wish to acknowledge at the outset that I make disproportionate reference to the work of myself and my collaborators in describing various techniques. I do so merely because so much of my own thinking is based on lessons from my own research and I can use specific examples from that work to vividly and concretely illustrate approaches that I believe have general applicability.

interventions. Third, a field experiment is designed. Fourth, the experiment is implemented and data is collected. Fifth, data is analyzed. Lastly, results are interpreted.

Topics

In selecting a specific topic for study, there will always be value in working in areas where there is interest in taking action based on results, and in areas that are understudied. Sometimes it will make sense to go directly from idea to field experiment. As complementary approaches, ideas for the topics of study for field experiments may be generated from earlier stages involving exploratory research methods that can be implemented more quickly and less expensively. The value of field experiments can be enhanced by imbedding them in a program of related research, so that the experimental work can interact synergistically with research of scholars involved and with the broader literature in order to build a deep, coherent, synthesized body of knowledge. This section illustrates the methodological technique of building field experiments at early stages that use two different approaches: non-experimental research and laboratory research.

Building on non-experimental research. Analyses of participation in a Tax Deferred Account offer an example of entrepreneurship in collecting new data and of creating a program of research building directly on earlier non-experimental research. This line of research focused on analysis of the decision to enroll in a savings plan by university staff, and an attempt was made to study the spillovers from colleagues in the same department on individual decisions. Duflo and Saez (2002) used an instrumental variables strategy for identification, assuming that average wage or tenure in a department did not directly affect individual decisions (conditional on individual wages or tenure) in order to use average wages or tenure as instruments for average

participation. They later followed up on this initial work to implement an experiment in which some individuals in some departments received incentives to visit a benefit fair to learn more about the savings plan. The effects of spillovers could be assessed by comparing the outcomes of the non-incentivized individuals in incentivized departments to individuals in non-incentivized departments. In analysis of the experiment, Duflo and Saez (2003) found that the estimates of the social effect were smaller in the experiment than in the preceding work using instrumental variables, but that small changes in the environment induced by the experiment did have a noticeable impact on plan participation.

Building on laboratory research. Studies of matching and rebate subsidies for charitable contributions have been the subject of a sustained program of research in both the laboratory and the field. The U.S. tax code provides a form of a rebate for charitable contributions by making them tax-deductible. As an example of economic equivalence, a 20 percent rebate on $1 of contribution means that the gift would cost the donor 80 cents. An 80 cent donation with a 25 percent match would also cost the donor 80 cents and raise $1 for the charity. The economically equivalent rebate may not be as successful as a match in encouraging charitable giving – perhaps because the rebate is more complicated, less certain to be given, or just that the rate associated with the rebate (20 percent) is lower than the match (25 percent). Eckel and Grossman built on their earlier laboratory experiments (2005) examining this issue by conducting field experiments (2006a, 2006b) with two different charities in Minnesota where potential donors were offered rebates or matches. Although each experiment had complications, the preponderance of the evidence was that the matches consistently raised more funds for the charities in both the laboratory and field experiments. Thus, it may be possible for the government to implement a

match received directly by charities having the same overall budgetary cost as the current deductions, but resulting in greater donations.[5]

Summary. A good topic for a field experiment can sometimes be identified from studies that may not be completely convincing but that offer intriguing results. One illustration is of non-experimental instrumental variables analysis of informational spillovers in saving decisions followed by a field experiment. Another is of a series of laboratory experiments on matching versus rebate incentives in charitable giving leading to two field experiments.

Conceptual framework

Economic theory provides a powerful set of tools to generate hypotheses about the potential impacts of policy interventions. In order to make the framework in which we consider the potential impacts of a policy intervention more complete, it can be helpful to look outside the toolbox of conventional economic theory. Two methods for doing so that have led to fruitful insights are the consideration of psychological factors and the development of hypotheses based on direct interaction with the individuals who will be affected by the intervention.

Consideration of psychological factors. There are various decisions relevant to public finance in which models of purely financial terms have been extended to include psychological factors, such as the potential importance of social stigma in the decision to take-up government benefits. Contributions to public goods have been of long-standing interest in public finance, and there have recently been a number of field experiments that examined psychological factors affecting giving. Field experiments can be very useful in establishing the magnitude of substantive importance to ascribe to psychological factors in decision-making by making direct

[5] The responsiveness of very high income individuals to the match would be a key factor in such calculations, and has yet to be explored.

comparisons to other factors in decisions such as prices. Because the specifics of context are so important in psychological factors, the clear incorporation of context within an experimental study of an intervention can enhance the value of knowledge generated by the experiment. To the extent that government subsidizes charitable giving through the tax code and may undertake additional policy interventions in the future to encourage giving, this line of work may also become relevant to public policy.

In one field experiment embedded in a general population survey, participants were asked to decide how they will split $100 between themselves and the local chapters of Habitat for Humanity in a small city that was heavily affected by Hurricane Katrina (with a 10% chance their decision will be implemented), with experimental conditions changing the information available about the residents of the city. Fong and Luttmer (2006) found that only income had a substantial effect on giving, which was more generous when Katrina victims are perceived to be poorer, while information about race had no effect.

Another field experiment was designed around university fundraising. List and Lucking-Reily (2002) found solicitations that described the current level of seed funding as 67 percent of the fundraising goal generated nearly six times more in gifts than seed funding of 10 percent. Offering to refund contributions if a goal was not met increased contributions by about 20 percent. Other work in this literature finds that donors are sensitive to expectations about the giving of other donors (Frey and Meier 2004; Shang and Croson 2007).

Direct interaction with potentially affected individuals. Anthropologists and sociologists often incorporate qualitative methods into their research that are based on finding patterns in the richer data of open-ended interviews and field observation. These methods can be incorporated into the process of creating a field experiment. The result is a mixture of methods for a larger

project including both open-ended exploratory interviews and closed-end survey questions. These are most often used iteratively in order to develop hypotheses based on both theory and context given from qualitative work, to test hypotheses quantitatively, to examine puzzling findings qualitatively, and then develop new hypotheses.

One example of hypothesis generation is a study of the use of housing vouchers to move out of neighborhoods with high concentrations of poverty. Kling, Liebman, and Katz (2005) reported that they had initially intended to focus on housing, labor market, and schooling effects of this intervention. However, an initial period of fieldwork visiting intervention facilities and conducting open-ended interviews with a small number of families resulted in a re-orientation of research to include health issues like depression and asthma in addition to the issues of employment and education that had been the original focus. In a follow-up survey, families offered housing vouchers by lottery to move to low-poverty neighborhoods had adults with better mental health, and children with fewer asthma attacks than in a control group not offered vouchers (Katz, Kling, and Liebman, 2001).

Summary. Hypotheses about the potential effects of interventions can come from economic theory, but can also come from consideration of psychological factors and from direct interaction with potentially affected individuals. In a series of charitable giving experiments, the body of evidence indicates that the framing of the conditions of recipients does influence giving levels. In a housing voucher study, open-ended interviews and field observation led to the hypotheses about impacts on health that suggested additional data to be collected.

Design

The design phase of an experiment is the unique aspect of the research. In contrast with correlational or quasi-experimental analysis, the researcher has the flexibility and the responsibility to create variation in the data that will plausibly identify a causal effect of an intervention rather than having to rely on unplanned variation that may not be plausibly related to the intervention of interest. There are many aspects of design, of which I will focus on four. First, the circumstances of individuals who do not receive the intervention (that is, the counterfactual) should be carefully considered. Second, any variations in the intervention should be chosen to provide the most information about the questions of interest. Third, the outcomes of interest need well-defined measures. Fourth, summary indices of key outcomes should be identified at the design stage to increase statistical power in analyses confirming key hypotheses.

Counterfactuals. The field experiment will be informative about the effect of the intervention relative to the experience of the control group. This makes the circumstances of the control group key to the policy relevance of the results. Typically, the control experiences whatever would have happened in the absence of the experimental intervention (that is, business as usual). Information on these circumstances is often available in the design stage, and is critically important to examine.

New Chance was a program that provided intensive employment services, on-site child care, and other services to young unmarried mothers. The program had few effects on either mother or child outcomes relative to control group (Quint, Bos, and Polit, 1997). It turned out, however, that the control group was able to tap into similar types of services in the community as those in New Chance and as a result the differences in the experiences of the intervention and

control groups were diminished. The study was a weak test of the effect of the intervention relative to its absence.

Intervention variations. In the social experiments of the 1970s, there was a conscious attempt to design studies that could provide estimates of a response function, such as the response of medical expenditure to the price of health insurance or labor supply to earnings subsidies, over ranges of price or income. These complicated designs offered many variations of the intervention, with small numbers of individuals receiving each variation. They were more susceptible to implementation problems and more difficult to explain to policy-makers in simple and understandable terms. In some cases the implementation problems involved with having many intervention variations can surely be solved, such as when the intervention is delivered entirely by computer. However, many interventions involve delivery of services by people, and people can be confused by multiple variations of an intervention, degrading the intervention itself. Reflecting on the experiences of early experiments, Hausman and Wise (1985) explicitly advocated simpler "black box" designs with sometimes only intervention being compared to a control group. Yet, use of more than one intervention can substantially enhance the value of the experiment.

One example of multiple interventions in a field experiment is that of welfare reform in Minnesota. One variant of the intervention involved only financial incentives, letting individuals keep more of their welfare benefit when earnings increased. Another variant combined financial incentives with work requirements, requiring longer-term public assistance recipients to work or participate in employment services. A control group had neither enhanced financial incentives nor work requirements. Gennetian, Miller, and Smith (2005) found that the financial incentives caused employment rates to increase somewhat and led to increased income on average, but also

led some people to reduce their hours of work. The resulting effects on earnings were offsetting and similar on average to that of the control group. Combining the financial incentives with work requirements did effectively increase both employment and earnings, demonstrating the additional effect of the work requirement – which would have been impossible to separate out without including the financial incentives only group in the experimental design.

Another example of multiple interventions involves a charitable giving experiment in which three groups that were randomly assigned to receive different offers of matching rates for donations relative to a control group. Karlan and List (2006) found that larger match ratios of $3:$1 and $2:$1 had no additional impact relative to match ratios of $1:$1. These results suggest that future research on policy interventions to encourage charitable giving with incentives should attempt to incorporate tests for nonlinearities in responsiveness to these incentives.

Well-defined measures. One of the most challenging elements of experimental design is determining the precise definitions of the outcome data to be collected. One may have a general idea of studying asset accumulation, for example. The challenge lies in translating this idea into specific data elements to be collected on home ownership and different types of financial assets such that they can be collected efficiently, cleanly aggregated and compared to national data, requiring attention to detail. Making trade-offs between what should be collected versus omitted given the constraints of survey data collection requires a strong conceptual framework to guide prioritization of measure selection.

To give one example of measure selection, a random assignment field experiment was used to study the impact of individual development accounts (IDAs) on asset accumulation. IDAs are saving accounts that provide low-income households with matching payments when the balances are withdrawn and used for special purposes. The IDA program examined matched

funds used for home purchases at 2:1 and funds used for business start-up and investment in education at 1:1. It also provided participants with financial education and counseling, as well as reminders and encouragement to make regular contributions. Survey data was collected at baseline, 18 months, and 48 months.

In developing measures of financial net worth, about which it is often difficult to obtain accurate reports, special effort was made to develop criteria to help identify and verify responses that might have been misreported or misrecorded. Responses were verified if they fell outside a specified range for each question, the change in the recorded value between one wave and the next fell outside a specified range, or the value was inconsistent with another response in the same wave. The results reported by Mills *et al.* (2006) show little overall effect on asset accumulation of IDAs. For the subgroup of African-Americans who were initially renters, the IDA raised home ownership rates by almost 10 percentage points over 4 years but reduced financial assets and business ownership.

Indices. We often want to know the answer to a broad question, such as whether an intervention improved economic self-sufficiency. Various measures, such as employment, earnings, lack of government assistance, and amount of government assistance all might contribute to our assessment that economic self-sufficiency was higher or lower. The relative weight given to different measures will inevitably be somewhat arbitrary. It is often simplest and most transparent to divide each measure by its standard deviation and average these together to form an index. Alternatives include principal components analysis, factor analysis, or least squares projection (say, from a regression using an auxiliary panel dataset to use multiple indicators to predict long-run average income). All will normalize the measures to a common scale and form some weighted average. The resulting index can later be used as an outcome in

experimental analysis to come to a general conclusion about the impact of an intervention on a broad domain such as economic self-sufficiency.[6]

It is valuable to specify indices at the design stage because doing so allows the analyst to later make a clear statement of what was to be tested and what the results were. When key indices are not specified in advance, then there are many results that could be reported and it is not clear how to choose which should be emphasized. If there was no true effect of the intervention but there were 100 intervention effect estimates, then we would expect 5 with t-statistics absolute values of 1.96 or greater just by chance. If the 5 most significant results from a set of 100 had t-statistics around 2 and were emphasized in the analysis, it would be hard to know if these simply occurred by chance. If there is one pre-specified index, then we can precisely characterize the probability that a result that large would occur by chance if the null hypothesis of no effect were true. It can also be used in the power calculations needed to assess whether the sample size will be sufficient to detect effects of plausible magnitude.[7]

The approach of using summary indices has been applied in re-analysis of the long-term effects of pre-school programs. Anderson (2007) created summary index measures for each of three programs for three parts of the life course (pre-teen, teen, and adult). The indices contain multiple measurements of variables such as IQ and educational attainment. Analysis focused on this set of nine outcomes. Analyses were conducted separately by gender to answer two separate

[6] At the analysis stage, the test statistic for intervention effects on this index will be most powerful against an alternative hypothesis that intervention impacts on all components of the index were equally beneficial, which is often the working alternative hypothesis in assessment of policy interventions. An F-test on intervention effects for each separate outcome included in an index would be a more powerful test against a non-directional alternative where some impacts might be beneficial and some adverse.

[7] Even if a single measure and an index are both normalized to have a standard deviation of one, the index will typically have a higher signal to noise ratio, which increases the statistical power. Note that a common error in power calculations involves setting the sample size such that the t-statistic will equal 1.96 for particular intervention effect size. If the true effect is of this size, then half the time the estimated effect will be higher half the time and lower half the time. Thus, the power to detect a true effect is only 50 percent when using this critical value. Conventional calculations usually involve 80 percent power and use a critical value of 2.8. See Orr (1999) for details.

questions: what is the evidence on the effectiveness of any one of these interventions at a life-course stage for females, and what is the evidence for males? The results show that females display significant long-term effects from early intervention, while males show weaker and inconsistent effects. Among the components of the indices, females receiving pre-school interventions show particularly sharp increases in high school graduation and college attendance rates, but they also demonstrate positive effects for better economic outcomes, less criminal behavior, less drug use, and higher marriage rates.

Summary. Design gives a field experiment unique flexibility to tailor variation in the data to answer the research question of interest. When the circumstances of the control group are known, the intervention can be designed so that there is a sufficient difference in experience between the control group and the intervention group. Without attention paid to the counterfactual, a well-designed experiment in every other way can yield null findings, similar to what happened in the study of services for young mothers. Planned variations in interventions can show the effects of components of a composite intervention (such as financial incentives and work requirements in welfare reform) or increased intensity (such as higher matching rates in a charitable giving experiment). Carefully selecting measures at the design stage translate a concept into something measurable (such as translating asset accumulation into housing and financial assets). Indices can aggregate information, form the basis of general conclusions, and protect against later distortion of statistical inference when specified at the design stage (as in analysis of the effect of early childhood interventions on later educational outcomes).

Data collection

In many areas of public finance, the data relating to a current or potential policy needs to be created for the research. As a point of departure, the thirty-year old advice of Rivlin (1974) still rings true today:

> Economists and others who seek to improve the basis for policy formulation should devote far more attention and resources to data collection than they have in the past. Ever fancier statistical manipulation of data from traditional sources is unlikely to improve policy choices significantly. The effort required to design and to carry out both surveys and experiments may be painful and unglamorous, but it is the *sine qua non* of more rational and informed policy formulation. (p. 353)

Seldom are our data measured without error, but key questions for the use of experimental data are whether there is a systematic bias in the data, and whether that bias differs across intervention groups. This section describes methods to assess potential biases. First, analyses of administrative records data provide an opportunity to check for the presence of bias. Second, combined analyses of administrative and survey data can reveal directions of bias.

Administrative data. A particular challenge for early social experiments was the collection of follow-up data from the control group, which had less incentive to stay in touch with program operators and which was harder to locate later on. In field experiments, the greatest single threat to the analysis is attrition. One way to minimize these potential problems is to focus on outcomes that are collected for everyone automatically through a record keeping system that collects data for an administrative purpose, typically covering a much broader population than individuals in the experiment.

An example using this approach was a field experiment designed to analyze the effects of matching rates on contributions to retirement savings accounts among low and middle income individuals. The experiment was implemented as an added incentive on top of the existing tax code in order to assess the potential impact of matching incentives that have been contemplated

but not implemented as policy interventions. Clients at a tax preparation firm were randomly assigned to groups offered either no match, a 20 percent match, or a 50 percent match if they opened an IRA at the time of their meeting with the tax professional. The offers of matches were added into an existing software system that collected data on the opening of IRAs. Although other outcome data such as net worth would certainly also be of interest, the built-in administrative data collection allowed the analysts to have complete data on account openings and later balances for everyone in both the intervention and control groups. Duflo *et al.* (2006) used these data for analysis and found a significant increase in retirement contributions in response to the matching offer, even among low and middle-income households with generally low propensities to save.

Survey and administrative data combinations. Since administrative records are by definition designed for administrative purposes other than research, they sometimes do not contain the detail that can be introduced into a survey created for research purposes. Moreover, administrative data may not have the geographical coverage needed or may miss certain important activities. Use of both survey and administrative data can make the results of experiments more informative.

Both types of data were used in the National Job Corps Study, which followed groups randomized into and out of receiving the Job Corps services of basic education, vocational skills training, health care and education, counseling, and residential support. According to both the survey and administrative records data analyzed by Schochet, McConnell, and Burghart (2003), the estimated earnings impacts are negative in the first and second years (the period during Job Corps enrollment) and positive and statistically significant in the third and fourth years after

random assignment. Assuming these earnings effects after four years would persist, cost-benefit analysis indicated that the social benefits of Job Corps were greater than the costs.

Comparison of results from the administrative and survey data found that reported earnings levels are much higher according to the survey data, perhaps because of incorrect social security numbers and employment not covered in the administrative records data or over-reporting of hours worked in the survey data. Also, earnings impacts were larger for survey respondents than for non-respondents according to the administrative records data, suggesting that the survey-based impact estimates are slightly biased upward. A key advantage of the administrative data was the ability to inexpensively extend the original analysis out to seven years after random assignment. In years five to seven, all the impacts were near zero. Revised cost-benefit estimates from 2003, based on more extensive data than those published in 2001, suggested that the benefits to society of Job Corps are smaller than the substantial program costs.

Another illustration of the value of using both survey and administrative data is taken from an analysis of the effect of housing vouchers and residential location on criminal behavior of youth. Kling, Ludwig and Katz (2005) found that female youth in the voucher groups had fewer property crime arrests, while the male youth in the voucher groups had more property crime arrests than the control group. One concern with administrative arrest data, however, was that it might be affected by differing policing intensity in different areas. If the probability of arrest for property crime conditional on offense is higher in low-poverty neighborhoods, then the appearance of an effect could be spurious. Another source of information was survey self-reports about behavior problems. These reports have their own measurement issues, but one virtue is that policing intensity likely plays a minor role in self-reports about trouble getting along with teachers, etc. The results based on self-reported behavior problems also showed an adverse effect

for male youth that was opposite in sign and significantly different from the effect on female youth.

Summary. Incomplete data could lead to biased estimates based on a field experiment, but administrative sources can provide complete data on some outcomes. Administrative data on IRA account openings was used to study the effects of matching incentives on savings. Survey and administrative data can be combined to assess potential bias. Both a Job Corps and a housing voucher study used survey data to assess reporting biases in outcomes levels based on administrative data and biases having different sources in survey data. These studies both found similar results for differences between intervention and control groups using the two types of data having two different types of errors, strengthening the cases that the estimates of the intervention effects were less likely to be artifacts of one particular type of reporting error. The use of multiple data sources from different perspectives greatly aided the formulation of a consistent set of substantive conclusions.

Analysis

The core of experimental analysis is the difference in average outcomes between intervention and control groups. This elegantly simple approach is straightforward to conduct and easy to understand. Complications for statistical inference arise when there are multiple comparisons of interest, which is the subject of the first part of this section. The second part of this section, addresses the question about how to predict what would happen if something similar but not quite the same were implemented. For forecasting these effects of untried policy interventions, a structural model can be used.

Inference from multiple comparisons. Assume you were interested in the cost-effectiveness of government funding for a Big Brother & Big Sister program, which may be expanded if it is successful for either boys or for girls. Given two separate intervention estimates (one for boys and one for girls), the probability both have t-statistics less than 1.96 when the true parameters are both zero is $.95^2 = .9025$ (under independence), so the probability one or both are 1.96 or higher is .0975. If we looked at ten age groups, the probability when the null hypothesis of no effect is true that at least one group will have a t-statistic of 1.96 or higher is $1-.95^{10} = .40$. The probability of observing a t-statistic around 2 is quite high even when there is no true effect.

The underlying problem is that when we engage in a process of searching across many comparisons for those that have particular value of a t-statistic, our inferences based on that t-statistic are distorted by the search process itself. Yet, it may be a core question of the research to know whether the intervention is effective for boys or for girls, for outcomes of educational achievement or delinquency. These questions involve sets of multiple comparisons. Fortunately, we can adjust our statistical inference to answer the question: what is the probability that the highest t-statistic observed in a set of comparisons would have occurred by chance under the joint null hypotheses that all of the true parameters in a set were zero? These probabilities are known as adjusted p-values. The adjusted p-value for a t-statistic of 1.96 in a set of two independent comparisons is .0975. When the estimates for two comparisons are likely to be correlated (say, as in the case of the education and delinquency outcomes), then an estimate based on an independence assumption is too conservative. A corrected estimate can be obtained using a bootstrap technique.

An illustration of special attention to statistical inference with multiple comparisons is the analysis the effects of housing vouchers on adult outcomes by Kling, Liebman, and Katz (2007).

They focused comparison of two intervention groups to a control group for four summary indices: economic self-sufficiency, mental health, physical health, and an overall measure. In this set of eight estimates, the effect on mental health had by far the largest effect size, and it had a t-statistic of 2.8. The adjusted p-value was .06 for the t-statistic on the mental health index being 2.8 or higher under the joint null hypothesis of no effect on any of the eight adult indices being considered.

Using the technique of adjusting p-values naturally leads to the question of which p-values should be adjusted. The authors pre-specified a set of eight key comparisons, and implemented their adjustments on this set. That means for these eight comparisons, they can precisely calculate the adjusted p-value. They kept the set limited for a reason, however. As the size of the set increases, the adjustments become larger and true effects become harder to detect. Despite restricting use of adjusted p-values to this key pre-specified set, the authors did report plenty of other comparisons. However, these other comparisons were discussed as being more exploratory and statistical inference about these other comparisons was not as rigorous.

For example, another notable result was that those who used housing vouchers to move were eleven percentage points less likely to be obese, with a t-statistic of 2.2. While obesity was a component of the physical health index, this comparison was not part of the set of primary outcomes. The authors looked at this result in two ways, depending on the process the reader used to get to that estimate. If an obesity researcher went straight to that point estimate because she was looking for experimental evidence of the effects of housing vouchers and residential location on obesity and was interested in this estimate regardless of its t-statistic because of the unique nature of the experiment, then traditional statistical inference would be valid as though this was the only outcome from the experiment (since for her, it essentially was). However, if a

public finance economist searched through the full range of results from the experiment and focused on the obesity result because of its high t-statistic, then she should be aware that the obesity result was more likely to have been caused by sampling variation than a non-adjusted p-value would suggest. Moreover, the fact that there were not impacts on other physical health measures that some theories suggested would have been affected in the same way as obesity suggested that the obesity result should be treated with caution.

Structural modeling. In order to base out-of-sample projections about policy alternatives on sound parameter estimates, empirical work from field experiments can be combined with general equilibrium modeling. If the structural model is specified during the design phase of the experiment, alterations to the design may even be suggested (and weighed against other factors) that would improve the usefulness of the model.

In examination of welfare reform policies that subsidize employment search, one group of researchers used data from Canadian welfare reform. This experimental intervention provided monthly cash payments for up to three years to long-term recipients of income assistance contingent on their finding full-time employment within one year of the supplement offer and leaving the income assistance program. The approach started with a textbook Pissarides matching model. The model was calibrated to the control group, and found to do a good job predicting the partial equilibrium outcomes of the single parents receiving the earning supplement intervention (Lise, Seitz, and Smith, 2005b). This was a much more stringent test of the model than the typical replication of general economic conditions, as it had to capture the effects of a policy change without directly using data on that change in the calibration.

Lise, Seitz, and Smith (2005a) went on to use the model to assess general equilibrium effects of the intervention. Among the key results were that employment increases for recipients

were offset by employment decreases among other groups for no net employment changes, and that subsidy recipients had higher exit rates from unemployment but received lower equilibrium wages. In contrast to partial equilibrium analysis, the general equilibrium analysis suggested that the benefits of the intervention did not outweigh the costs.

Summary. First, when examining multiple comparisons in an experiment, the probability of observing a t-statistic of 1.96 or higher rises with the number of comparisons under the joint null hypothesis of no effect. Adjusted p-values can be used to rigorously show the probability that highest t-statistic in a set would be at least that high if there were no effect. This technique is valuable when a set of estimates for key pre-determined outcomes are all important, and one wants to make statistical inferences about the most significant estimates in the set – as illustrated in statistical inference about adult mental health effects in a housing voucher experiment. Second, use of a structural model provides a way to simulate a much broader range of policy alternatives while remaining grounded on parameter estimates in which there is high confidence and rigorously benchmarked against particular interventions for which the results have been observed, as illustrated for a welfare reform experiment.

Interpretation

As soon as the impacts (or lack of impacts) from an intervention are clear, the question of why those effects occurred immediately follows. This interpretative process is the phase of the experiment that moves beyond the results of the experiment on the primary outcomes. For understanding results, first a detailed description is needed of how the experiment was implemented and sustained (or, not) during the follow-up. Second, non-experimental analyses may help draw linkages between particular aspects of an intervention and subsequent effects on

outcomes. Third, narratives from open-ended interviews can be studied for patterns that suggest hypotheses consistent with the observed results that can be tested in future work.

Intervention description. Understanding interventions can be complicated. It is helpful to understand the variability in the intervention across locations and across time. Sometimes interventions have multiple components. Even when the intervention itself is simple, the intermediate pathways that it may affect an outcome of interest can be complicated. For example, let's say we were interested in the effect of a housing voucher on the educational achievement of children. The housing voucher could affect the housing quality, the neighborhood poverty and safety, the school quality, parenting practices, and a host of other factors. Address data from program operations and survey data can help describe both what the circumstances of intervention group were and how they differed from the control group. Sanbonmatsu *et al.* (2006) found that the intervention caused sharp changes in neighborhood poverty and safety, but relatively little change in school quality and essentially no change in parenting. The fact that there were no discernable impacts on school achievement suggested a logical argument that the linkage between neighborhood factors and school achievement was not strong. Since the intervention had little impact on school or parents, the effects of these factors would likely be better investigated using another experimental design.

Non-experimental analysis. There are a variety of ways to use non-experimental methods to learn from data collected from an experiment. For example, it is often quite useful to examine the predictors of who participates in an intervention when it is offered to them. In this section, I focus on one particular type of non-experimental analysis, combining planned variation across intervention types and unplanned variation across sites in experiments that are conducted in multiple locations. As with the description of the intervention, the goal of these analyses is to

look inside the black box of an intervention. However, instead of using a chain of logic to suggest that a relationship does or does not exist, the method used here relies on the formulation of an econometric model and its attendant assumptions.

One example of the use of such a model was in analysis of welfare reform involving two interventions at each of three sites. Welfare recipients with young children were randomly assigned to either an education-focused group, a work-focused group or to a control group that received no additional assistance. In order to examine the responsiveness of child outcomes to changes in the mother's employment and education, the variation in treatment effects across interventions and sites was projected onto the two dimensions of maternal employment duration and maternal education level using a two stage least squares instrumental variable estimation.

Estimates based on this method are valid if the assumption holds that the effect of the intervention works through the two hypothesized mechanisms and not through other omitted variables. Since the variation across sites was not planned into the experimental design, the site differences in intervention effects could be related to some other factor (such as local economic conditions), but sometimes a convincing case can be made that the econometric model has captured the factors of primary importance. Magnuson (2003) found that increases in maternal education are positively associated with children's academic school readiness, and negatively associated with mothers' reports of their children's academic problems, but with little to no effect on children's behavior. This methodology and related extensions are described in Gennetian, Magnuson, and Morris (2006).

In a second example of this technique, a control group and treatment groups receiving two different types of housing vouchers across five different sites allowed use of treatment-site interactions to instrument for measures of neighborhood crime rates, poverty and racial

segregation in analysis of which neighborhood factors affect individual arrest outcomes. The underlying logic of this approach is that interventions at some sites had large impacts on area crime and not area poverty while other sites had larger impacts on area poverty than area crime. A useful device for putting magnitudes of effects in context was to scale measurements in standard deviation units so that a large change in neighborhood poverty rate could be meaningfully compared to a large change in neighborhood crime rate. While it is not likely to be true that any two particular characteristics of neighborhoods are the only ones through which a housing voucher intervention works, one would have expected bias in estimates of the effect of local area crime on own crime to result in estimates that were large and positive. The lack of an effect found by Ludwig and Kling (2007) even when one might have expected upward bias indicated lack of support for the hypothesis that local area crime increases own crime. Neighborhood racial segregation appears to have been the most important explanation for variation across neighborhoods in arrests for violent crimes, perhaps because drug market activity is more common in high-minority neighborhoods.

Patterns in narrative data. Using survey data collected for this same housing voucher study, a surprising pattern of results was found in which teen girls appeared to have benefited from being in a family encouraged to move to a lower-poverty neighborhood with a housing voucher while teen boys were affected adversely. Clampet-Lundquist *et al.* (2006) showed these gender differences in effects, and then examined patterns in the narratives from open-ended interviews to generate several new hypotheses. Teen boys appeared more isolated from the mainstream when they moved to areas of lower poverty. Teen boys in new neighborhoods tended to spend their free time on neighborhood basketball courts and on street corners, in closer proximity to illegal activity than the teen girls who tended to spend free time at home or in malls

or other more supervised spaces. Teen boys in families offered housing vouchers also had relatively lower rates of contact with father figures (stepfathers, mother's boyfriends, and uncles). All of these factors may have contributed to the gender differences in effects.

Another example of mixing narrative data analysis in study of a policy experiment is the examination of a bundle of anti-poverty services known as the New Hope intervention. New Hope required proof of 30 hours of work per week, and provided an earnings supplement that raised income above the poverty line, subsidized child care, subsidized health insurance, and a temporary community-service job if needed. Duncan, Huston, and Weisner (2007) found that poverty rates declined dramatically, and that employment and earnings increased among participants who were not initially working full-time. Among children in families receiving the intervention, school performance improved (especially for boys) with academic gains equivalent to half of the average achievement gap between black and white school kindergartners.

One of the important and initially puzzling findings from New Hope was on teacher-reported achievement and behavior of younger children. In the experimental group boys, but not girls, were 0.3 to 0.5 standard deviations better behaved and higher achieving than their control-group counterparts. Based on the survey data alone, however, the authors were unable to understand this gender difference. A systematic analysis of qualitative data from 44 families, sampled randomly from both program and control group members, focused on instances where parents made specific references to the gender of their children, shed light on this puzzle (Duncan and Gibson, 2005). Respondent narratives made clear that mothers believed that gangs and other neighborhood peer pressures were much more threatening to their boys than their girls, and that mothers in the experimental group channeled more of the program's resources (e.g., child care subsidies for extended-day programs) to their boys. One mother put it this way: "Not

all places have gangs, but [my neighborhood] is infested with gangs and drugs and violence. My son, I worry about him. He may be veering in the wrong direction...it's different for girls. For boys, it's dangerous. [Gangs are] full of older men who want these young ones to do their dirty work. And they'll buy them things and give them money."

Summary. The interpretation of results extends beyond discussion of the effect of the experimentally tested intervention on outcomes. Descriptive information about the implementation of the intervention provides details about what the conditions that comprised the intervention and affected participants – for example, characterizing the neighborhoods and schools of youth in a housing voucher study. Intervention variants and locations can serve as instrumental variables for selected conditions within the intervention that can be linked to their effects on outcomes, as demonstrated in a welfare reform study and a housing voucher study. Patterns in narrative data generate suggestive new hypotheses that potentially explain why certain intervention effects may have been observed and that can be tested in future research through collection of larger samples of data.

Conclusion

Although this article is organized around the linear end-to-end process of conducting a field experiment, there are two main cross-cutting themes that are important in multiple phases of the research. First, greater statistical sophistication can draw more value from a field experiment without obscuring the simple and compelling information from the differences in average outcomes of intervention and control groups. Non-experimental econometric techniques can form the basis of studies that motivate the topics of field experiments, and that extend the analysis to look inside the black box of the experiment. Creation of indices and focus on key

outcomes during design can lead to more credible statistical inference, and these inferences can also be made more meaningful in studies where there are multiple comparisons of interest. Structural models can be calibrated to policy experiments and then used to simulate policy variations.

Second, the methodological frontier is interdisciplinary. Hypotheses are generated from consideration of psychological factors. Direct interaction with participants both generates hypotheses and assists in developing a conceptual framework rooted in the precise nature of the intervention. The interpretation of the results is enhanced by observation of patterns in narrative data relating those in survey and administrative data. These activities draw on knowledge and techniques developed in psychology, anthropology, and sociology that can be adapted for use in making public finance field experiments more useful.

Perhaps an economics graduate student starting her public finance dissertation today will think big, use field experiments, and incorporate the methods suggested here into her research. Indeed, the use of field experimentation itself in academic public finance appears to be growing.[8] Acquiring the methodological expertise to reach and to advance the frontier of experimental field research in public finance will require the intellectual plasticity to absorb new ways of thinking. More generally, use of these methods will help engage the creative and vibrant work that will produce results that are practical and credible. Most importantly, widespread adoption of methods suggested here has the potential to make the results from field experiments more useful to public finance and to public policy debate.

[8] For example, Greenberg, Shroder, and Onstott (1999) reported that the share of new ongoing social experiments led by academics was about 50 percent higher than the share of projects led by academics over the previous 34 years.

References

Anderson, Michael. "Multiple Inference and Gender Differences in the Effects of Preschool: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." Unpublished manuscript, University of California-Berkeley, 2007.

Burtless, Gary T. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives* 9 No. 2, (Spring 1995): 63-84.

Clampet-Lundquist, Susan, Kathryn Edin, Jeffrey R. Kling, and Greg J. Duncan. "Moving At-Risk Youth Out of High-Risk Neighborhoods: Why Girls Fare Better Then Boys." Princeton IRS Working Paper 509, March 2006.

Duflo, Esther, William G. Gale, Jeffrey B. Liebman, Peter R. Orszag, and Emmanuel Saez. "Saving Incentives for Low- and Middle-Income Families: Evidence from a Field Experiment with H&R Block." *Quarterly Journal of Economics* 121 No. 4 (November 2006): 1311-1346.

Duflo, Esther, and Emmanuel Saez. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment," *Quarterly Journal of Economics*, 118 No. 3 (August, 2003): 815-842.

Duflo, Esther, and Emmanuel Saez. "Participation and Investment Decisions in a Retirement Plan: the Influence of Colleagues' choices," *Journal of Public Economics* 85 No. 1 (July 2002): 121-148.

Duncan, Greg J. and Christina Gibson, "Qualitative/Quantitative Synergies in a Random-Assignment Program Evaluation." In *Discovering Successful Pathways in Children's Development: Mixed Methods in the Study of Childhood and Family Life*. Edited by Thomas S. Weisner. Chicago: The University of Chicago Press, 2005. 283-303.

Duncan, Greg J., Aletha C. Huston, and Thomas S. Weisner. Higher Ground: New Hope for the Working Poor and Their Children. New York: Russell Sage Foundation, 2007.

Eckel, Catherine and Phillip Grossman. "Subsidizing Charitable Giving with Rebates or Matching: Further Laboratory Evidence." *Southern Economic Journal* 72 No. 4 (April 2005): 794-807.

Eckel, Catherine and Phillip Grossman. "Subsidizing Charitable Contributions: A Field Test Comparing Matching and Rebate Subsidies." Unpublished manuscript, University of Texas-Dallas, 2006a.

Eckel, Catherine and Phillip Grossman. "Subsidizing Charitable Contributions in the Field: Evidence from a Non-Secular Charity." Unpublished manuscript, University of Texas-Dallas, 2006b.

Ferber, Robert and Werner Z. Hirsch. "Social Experimentation and Economic Policy: A Survey." *Journal of Economic Literature* 16 No. 4 (December, 1978): 1379-1414.

Frey, Bruno S. and Stephan S. Meier. "Social Comparison and Pro-Social Behavior: Testing Conditional Cooperation in a Field Experiment." *American Economic Review* 94 No. 5 (December, 2004):1717-22.

Fong, Christina and Erzo F. P. Luttmer. “Race and Giving to Hurricane Katrina Victims: Experimental Evidence.” Unpublished manuscript, Harvard University, 2006.

Gennetian, Lisa A., Katherine Magnuson, and Pamela Morris. “From Statistical Associations to Causation: What Developmentalists Can Learn from Instrumental Variables Techniques Coupled with Experimental Data.” Unpublished manuscript, MDRC, November 2006.

Gennetian, Lisa A., Cynthia Miller, and Jared Smith. *Turning Welfare into a Work Support: Six-Year Impacts on Parents and Children from the Minnesota Family Investment Program.* New York: MDRC, 2005.

Greenberg, David, Mark Shroder, and Matthew Onstott. “The Social Experiment Market.” *Journal of Economic Perspectives* 13 No. 3 (Summer, 1999): 157-172.

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

Harrison, Glenn W. and John A. List. “Field Experiments.” *Journal of Economic Literature* 42 No. 4 (December, 2004): 1009-1055.

Hausman, Jerry A. and David A. Wise. *Social Experimentation*. Chicago: University of Chicago Press, 1985.

Heckman, James, and Jeffrey Smith. “Assessing the Case for Social Experiments,” *Journal of Economic Perspectives* 9 No. 2 (Spring, 1995): 85-110.

Karlan, Dean and John A. List. “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment.” Unpublished manuscript, Yale University, 2006.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. “Moving To Opportunity In Boston: Early Results of a Randomized Mobility Experiment.” *Quarterly Journal of Economics* 116 No. 2 (May 2001): 607-654.

Kling, Jeffrey R., Jeffrey B. Liebman and Lawrence F. Katz. “Bullets Don't Got No Name: Consequences of Fear in the Ghetto.” In *Discovering Successful Pathways in Children's Development: Mixed Methods in the Study of Childhood and Family Life.* Edited by Thomas S. Weisner. Chicago: The University of Chicago Press, 2005. 243-281.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. “Experimental Analysis of Neighborhood Effects.” *Econometrica* 75 No. 1 (January, 2007): 83-119.

Kling, Jeffrey R., Jens Ludwig and Lawrence F. Katz. “Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment.” *Quarterly Journal of Economics*, 120 No. 1 (February, 2005): 87-130.

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. “Equilibrium Policy Experiments and the Evaluation of Social Programs.” Unpublished manuscript, University of Michigan, 2005a.

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. “Evaluating Search and Matching Models Using Experimental Data.” Unpublished manuscript, University of Michigan, 2005b.

List, John A. and David Lucking-Reiley. “The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign.” *Journal of Political Economy* 110 No. 1 (February, 2002): 215-233.

Ludwig, Jens, and Jeffrey R. Kling. "Is Crime Contagious?" *Journal of Law and Economics*, 2007, forthcoming.

Magnuson, Katherine. "The Effect of Increases in Welfare Mothers' Education on Their Young Children's Academic and Behavioral Outcomes: Evidence from the National Evaluation of Welfare-to-Work Strategies Child Outcomes Study." Institute for Research on Poverty Discussion Paper no. 1274-03, 2003.

Mills, Gregory, William G. Gale, Rhiannon Patterson, and Emil Apostolov. "What Do Individual Development Accounts Do? Evidence from a Controlled Experiment." Unpublished manuscript, The Brookings Institution, 2006.

Moffitt, Robert M. "The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs." *American Behavioral Scientist* 47 No. 5 (January 2004), 506-540.

Orr, Larry. *Social Experiments.* New York: Sage Publications, 1999.

Quint, Janet C., Johannes M. Bos, and Denise F. Polit. *New Chance: Final Report on a Comprehensive Program for Young Mothers in Poverty and Their Children*. New York: MDRC, 1997.

Rivlin, Alice. How Can Experiments Be More Useful?" *American Economic Review* 64 No. 2 (May, 1974): 346-354.

Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. "Neighborhoods and Academic Achievement: Results from the MTO Experiment." *Journal of Human Resources* 41 No. 4 (Fall 2006): 649-691.

Schochet, Peter Z., Sheena McConnell, and John Burghard. *National Job Corps Study: Findings Using Administrative Earnings Records Data - Final Report.* Princeton, NJ: Mathematica Policy Research, 2003.

Shang, Jen and Rachel Croson. "The Impact of Social Comparisons on Nonprofit Fundraising." *Research in Experimental Economics*, 2007, forthcoming.