SOCIAL EXPERIMENTATION, PROGRAM EVALUATION, AND PUBLIC POLICY

Maureen A. Pirog, Editor and JPAM Classics Series Editor

▼ Preface

▼ Introduction
Maureen A. Pirog, The State of Social Experimentation and Program Evaluation

▼ Social Experiments Versus Quasi-Experiments
SOCIAL EXPERIMENTATION,
PROGRAM EVALUATION,
AND PUBLIC POLICY

Maureen A. Pirog, Editor and JPAM Classics Series Editor

▼ Preface

▼ Introduction
Maureen A. Pirog, The State of Social Experimentation and Program Evaluation

▼ Social Experiments Versus Quasi-Experiments
Randomized Experiments


Quasi-Experiments
A. Natural Experiments


B. Pretests and Posttests with Comparison Groups and Selection Controls


C. Interrupted Time Series with Comparison Groups


D. Posttests Only with Comparison Groups


E. Regression Discontinuity Design
Meta-Analyses

Implementation, Performance Management, and Program Impacts

Ethics and Human Subjects
Howard Rolston. (2005). To learn or not to learn. JPAM, 24(4), 848–849.

The Use of Program Evaluations by Policy Makers
Dedication

For my loving parents, thank you for a lifetime of support and encouragement. And, for my children who love to live life in the tails of the distributions.
Preface

Program evaluations, including social experiments, comprise the lion's share of articles in the Journal of Policy Analysis and Management (JPAM), and these techniques hold a preeminent position in the field of policy analysis. Evaluation research asks the fundamental questions of interest to all evidence-based policymakers and administrators: What programs, policies, laws, and administrative innovations work? Do some work better than others, for whom, and under what conditions? Addressing the issues of what works or what works better is not always straightforward and involves fundamental research design topics including measurement, sampling, study design, estimation procedures, accurately portraying results, and drawing relevant, concrete policy recommendations from the evaluation or experiment.

While this list of topics might sound a tad dry to newcomers to the field, the art and science of social experimentation and program evaluation is exciting, evolving, and critically important. There is an ongoing, important, and lively debate about whether or not social experiments should be considered the "gold standard" for evaluative work. Do the time, expense, and additional effort required to undertake a randomized field experiment really translate into more reliable impact estimates? In theory, perhaps yes. But the need to obtain agreement from individuals, schools, or other units to participate in a randomized experiment can introduce differences between the population of interest for a program and the population that has agreed to randomization—limiting external validity. Randomization can run amok. Randomized experiments are carried out in political arenas, and many other factors can result in differential attrition from treatment and control groups, which can undermine randomized experiments, leaving them subject to many of the same criticisms leveled at quasi-experiments. And over the past decade, a great deal of work has been conducted on new techniques to control for selection bias, in large part to help place quasi-experiments on a more equal footing with randomized experiments. And if that were not enough, there are always new programs, laws, policies, and managerial innovations being proposed and enacted in a very dynamic, changing society.

As the director of a research institute that undertakes evaluation projects and as a professor who teaches program evaluation, assembling this volume has been a learning experience. By the end of this experience, I have come to believe that most of our program evaluation textbooks are outdated. First, while randomized experiments are considered the "gold standard" by many, they actually represent a minority of the published evaluation articles in JPAM and several other journals focused specifically on evaluation (although medical and prevention sciences researchers may be more likely to publish random assignment studies in their own journals). Second, you will not find many (or in some cases, any) posttest only, Solomon four-group, simple time series, or similar overly simplified evaluation designs in JPAM. In our evaluation textbooks these
designs simply serve the role of straw men. There are no general internal validity that these types of studies are not in place. The policy scholars. Third, a large portion of the published work is on “natural” experiments in which cities, states, countries test policies and researchers capitalize on this natural variation to evaluate programmatic impact. Typically, researchers construct an ingenuity—combing individual-level census or survey data with variables and with geo-coded neighborhood data or other information to help them find larger cross-national studies. Researchers will often use two or three or even more distinct sources to construct their model because “adopters” may be fundamentally different in different locations, the topic of self-selection bias takes on considerable importance.

So it is with enthusiasm as well as a sense of hope that I have assembled this volume of articles on the field of program evaluation. As the current editor-in-chief of JPA, I have had access to such a wealth of material for this volume. For every article that was included, there were many more that could have been included. Ultimately, it is my hope that practitioners, and researchers will share my view that this report represents the state of the discipline of program evaluation research and provides perspective that will help advance the field.

Editor-in-Chief, Journal of Policy Analysis and Management
Rudy Professor, School of Public Policy

* This volume was assembled while visiting the Daniel J. Evans School of Public Policy and Governance, University of Washington.

Journal of Policy Analysis and Management  DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management
The State of Social Experimentation and Program Evaluation

Maureen A. Pircabart

The issue of causality is at the heart of public program evaluation. Does some deliberate manipulation, sometimes called a “trial,” “treatment,” or “independent variable” cause a change in specified “outcomes” or “dependent variables” that are theoretically tied to one another? I take a fairly broad view of what constitutes a “treatment” in this volume. “Treatments” largely comprise government programs, policies, legislation, and/or public management innovations.

SOCIAL EXPERIMENTS VERSUS QUASI-EXPERIMENTS

How one arrives at a basis for comparison gives rise to the distinction in the program evaluation literature between experimental and quasi-experimental methods. Experimental studies, random assignment (for example, coin toss, die, page of random numbers, computerized random number generator) is used to assign treatments to units. Among many other possibilities, units could consist of individuals, classrooms or schools in education studies, or plots of land in agricultural or environmental studies. Random assignment is considered the “gold standard” in evaluative research simply because, ceteris paribus, units assigned to the treatment and control groups should be statistically equivalent on both observable and unobservable characteristics, providing the sample size is sufficiently large. In contrast, in quasi-experimental studies, some other nonrandom method of selection is used to assign treatments to units: first-come, first-served; neediness; perceived ability; and so on.

When nonrandom assignment processes are employed, units are likely to differ on measurable and unmeasurable characteristics. From an evaluation perspective, measurable differences (for example, age, race, weight, height, number of children, income) are not problematic as they can be controlled in statistical analyses, providing that they are reliably measured and fully represent any constructs which they purport to measure. However, unmeasurable differences (for example, intelligence, motivation, altruism) are much more problematic because if they are correlated with the outcome measures, they can produce sizable biases on program impact estimates. This source of bias is called selection bias, or self-selection bias if unobserved for the treatment.

Undertaking randomization requires designing an experiment prior to implementing the treatment so that the treatments and controls can be randomly assigned to units. Often treatment dosages or interactions between multiple treatments are of policy interest, and these factors also need to be taken into consideration prior to program implementation. Unfortunately, only a minority of evaluation studies are designed prior to program implementation. For evaluations that begin after the program, policies, laws, or innovations have already been implemented.
Experimental and quasi-experimental designs or less rigorous methods (including making no evaluation at all) are the only options.

It should also be noted that while the random assignment of treatments to units may seem straightforward, in practice, this process can be complex. The trick is to design a randomization process that cannot be undermined by program operators who may have incentives to demonstrate program effectiveness, often by "crawling" treatments; treatment units who may want to opt in or out of treatment; or politicians who may want to ensure that the treatment units in their political districts (example, persons, schools, hospitals, water treatment plants) receive what is termed a priori to be preferred. At times, implementing randomization may require modification of complex client eligibility determination software packages to ensure that randomly selected applicants are eligible only for the treatment or control group. Some experiments have explicitly incorporated the pressures to undermine the randomization process into their initial designs (see King et al., 2007).

Finally, it has been argued, although never definitively proven, that randomized experiments are more expensive to conduct relative to nonexperimental evaluations simply because of the costs associated with implementing and monitoring randomization. If this common wisdom is correct, then these additional costs must be weighed against the costs of misallocating societal resources due to selection making based on less reliable nonexperimental studies.

Despite this generally accepted view that randomized experiments are the "gold standard," only eight studies published in JPAM since 2003, less than 10 percent of overall evaluative studies published in JPAM, were randomized experiments. Moreover, randomized experiments comprised only a slightly higher percentage of all evaluations in journal articles in the evaluation review, although a review of evaluation articles in medical or profession sciences journals may well produce a considerably higher proportion of randomized experiments. It is in this context that readers should be keenly interested in the first section of this volume, which is a debate on the relative merits of experimental and quasi-experimental methods between Robinson Hollister and Richard Nathan.

Additionally, in a volume such as this, we cannot ignore the emergent literature on various methods for correcting for selection bias. James Heckman received the Nobel Prize in Economics in 2000 for his pioneering work on correcting for selection bias. Since then, a variety of other methods have been developed to correct for selectivity, including propensity score matching, difference-in-differences methods, treatment effect models, instrumental variable approaches, and random effects models. Each of these selection correction approaches has limitations and can only be used under certain circumstances. A review of manuscripts published since 2003 in JPAM reveals that a sizable majority of the experimental program and policy evaluations use one or more of these selection correction methods. Therefore, a natural question emerges: How good are the selection correction methods? This question is addressed in the first section of this volume by three articles that directly to this debate. The first, "Do Experimental and Nonexperimental Evaluations Give Different Answers about the Effectiveness of Government-Funded Programs?", uses meta-analysis to look at whether or not random assignment studies produced different conclusions about the efficacy of government-funded training programs than quasi-experimental studies. The authors, David H. Berg, Charles Michalopoulos, and Philip K. Robins, conclude that the two types of studies reach similar conclusions, although they find differences in the magnitudes of the effects for youth and men and differences in variation of impact estimates for women. The second article related to the debate on the relative merits of
experimental and quasi-experimental research is “How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment” by Elizabeth Ty Wilde and Robinson Hollister. Using several ways to assess “closeness,” these authors conclude that propensity score matching does not yield impact estimates that are “close enough” to experimental impact estimates. Yet another perspective is provided by Thomas D. Cook, William R. Shadish, and Vivian C. Wong in their article, “Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons.” The authors find that some studies are better able to reproduce experimental results: studies using the regression discontinuity design; careful intact group matching when the groups are geographically proximate and matched on the pretest; and studies in which the treatment and comparison groups differ at the pretest but the selection process can be accurately modeled.

RANDOMIZED EXPERIMENTS

The next section of this volume includes three randomized experiments. The topics were selected in the hopes that they would be of interest to students enrolled in program evaluation classes. These studies focus on the impacts of abstinence-only education on teen sexual behaviors and outcomes; anti-poverty programming effects on marriage; and the efficacy of the Teach for America program on student achievement. Instructors looking for randomized experiments for their classes should not construe this list to be exhaustive. There are many other randomized experiments published in JPAM. Here are a few more, simply to make it easier for instructors to select readings that are particularly relevant for their courses and the interests of their students.


QUASI-EXPERIMENTS

While randomized experiments typically require at least some original data collection, many (but obviously not all) quasi-experimental evaluations use secondary data already collected from the U.S. Census, the Survey of Income and Program Participation (SIPP), the National Longitudinal Survey of Youths (NLSY), the Panel Study of Income Dynamics (PSID), Fragile Families, state or city administrative databases, or other such sources. This is not to understate the considerable effort it takes researchers. Usually data are combined from multiple sources. For example, in Houston and Richardson's 2006 JPAM study on seat belt use and traffic fatalities, they extracted traffic fatality data from the National Highway and Traffic Safety Administration's (NHTSA) Fatal Reporting System (FARS) for the years 1990–2002. Additional data were obtained from the Federal Highway Administration, the Insurance Institute for Highway Safety, NHTSA's Traffic Safety Facts, the U.S. Census, and the U.S. Department of Labor. Similarly, Lopoo and DeLeire's 2006 study of the impact of welfare reform on teen fertility used natality reports from the National Center for Health Statistics, demographics from the U.S. Census data, birth data from the National Vital Statistics Reports, and state-level data from a variety of additional sources. These studies are not exceptional in this regard.

Because the credibility of the research depends heavily on the quality of the data used, this topic deserves some elaboration. While international evaluations appear increasingly in JPAM, they still constitute a rather modest proportion of the publications. Recent topics have included deforestation in Bolivia (Andersson & Gibson, 2007); public management reform in the U.K. (Walker & Boyne, 2006); pollution in Canada (Harrison & Antweiler, 2003); pollution, economic growth, and national debt in Latin America (Aubourg, Good, & Krutilla, 2008); traffic safety in the European Union (Albatlute, 2008); and Colombian sickness funds (Trujillo & McCalla, 2004). These studies, however, still rely on large, official sources of data and generally combine data from multiple sources. For example, the article on debt, democratization, and development (Aubourg, Good, & Krutilla, 2008) utilizes data on 29 Latin American and Caribbean countries from the 2001 World Bank Development Indicators CD-ROM, with data inconsistencies being completed or corrected by data published by Haiti's Central Bank, the Caribbean Center for Monetary Studies at the University of the West Indies in Trinidad and Tobago, and the Center for Latin American Monetary Studies in Mexico. The Albatlute (2008) study on blood alcohol content limits and traffic fatalities in the former EU15 countries relied on the European Community database on Accidents on the Roads in Europe, Eurostat, WHO Europe, World Bank Development Indicators, and World Road Statistics.

Because of the fundamental nature and generalizability of the data provided by the U.S. Census, they are used frequently in evaluation research. Recent studies exemplify the diversity of topics that can be addressed using Census as a primary or secondary data source: The impact of urban enterprise zones on manufacturing establishments (Greenbaum & Engberg, 2004); the effects of full-day kindergarten (Cannon, Jacknowitz, & Painter, 2006); the effects of contracting out on public sector employment (Fernandez, Smith, & Wenger, 2007); immigration and welfare reform (Haider et al., 2004); and the effect of state R&D tax credits on private sector R&D (Wu, 2005).

Similarly, state-level and city-level databases have proven to be a gold mine for longitudinal assessments of state programs and policies as well as interstate
comparison of public policies where there is variation in policies across states. State administrative databases have been particularly useful in education studies (Wilde & Hollister, 2007), evaluation of state children's health insurance programs (Marton, 2007), child support programs, and welfare reforms (Page, Spetz, & Millar, 2005; Loeb et al., 2003). Securing confidential data from state agencies can be challenging, as can the sheer size of many of these data sets. There were over 46,000 observations in Marton's 2007 study of Kentucky's children's health insurance program. Even city-level studies can produce large data sets. For example, when examining the impacts of subsidized rental housing on property values in New York City, data were amassed on 432,984 property sales (Ellen et al., 2007). In the same city, Joyce, Gibson, and Colman (2005) examined differences in birth weights for 811,190 singleton births comparing infant birth weights for women who were on and off of the Women, Infants, and Children (WIC) program.

To summarize, most evaluation studies published in JPAM are quasi-experimental and most, but obviously not all, of these studies rely on large secondary data sources. Many of these secondary data sources are international organizations, foreign governments, and large federal databases such as those constructed and maintained by the U.S. Census Bureau, the U.S. Department of Labor, the Environmental Protection Agency, and the Department of Justice. State and city agencies can also provide detailed data to researchers. Also, large special databases, usually individual and family surveys, are collected for the specific purpose of conducting behavioral and policy research. The relatively heavy reliance on secondary data sources for quasi-experiments may well reflect the disciplinary biases of JPAM authors, who are largely drawn from the fields of policy analysis, public management, economics, political science, sociology, social work, and demography.

A total of 11 quasi-experiments are included in this volume. I have divided these articles into categories that reflect their design. The first four articles are "natural" experiments in which researchers capitalize on variation in laws or policies across time and geographic areas. Natural experiments constitute an increasingly large portion of the quasi-experimental work found in JPAM.

Natural Experiments

Four natural experiments are included in this section. The topics include the effects of selective serotonin re-uptake inhibitors (anti-depressants) in 27 countries; the impact of lowering the blood alcohol content (BAC) level to 0.5 percent for the original EU15 countries; the impact of variation in state laws prior to the Family Medical Leave Act (FMLA), and variation in coverage across states post-FLMA on leave-taking, employment, and earnings; and the impact of state and local antidiscrimination policies on the earnings of gays and lesbians. These "natural" experiments capitalize on variation in laws and policies across some geographic or political units and time and rely on panel data.

The "natural" aspect of these studies is that it is the state, locality, or country that initiates the new law, program, or policy—not a researcher who is randomly or otherwise assigning treatments to units. These studies are not true experiments; however, the geographic or political units initiating the innovations—the leaders—may be different from followers or non-innovators in ways that can be measured, as well as ways that cannot be measured. This is why a rather high proportion of the recently published articles that fall into this category make use of one of the selection correction methods.

Three of these studies explicitly take into consideration the issue of selection bias. In their study of antidepressants and suicide, Ludwig and Marcotte (2005) rely on country fixed-effects and country-specific linear trends to capture differences over
time and countries. Albalate's study (2008) of blood-alcohol limits in the European Union utilizes the differences-in-differences method as well as fixed effects. The Waldfogel study (1999) of the FMLA relies on a difference-in-difference-in-difference methodology to address potential sources of bias.

**Pretests and Posttests with Comparison Groups and Selection Controls**

Two studies are included in this section. Cuellar, McReynolds, and Wasserman (2006) ask the question: Do intensive mental health services provided through the Special Needs Diversionary Program (SNPD) reduce criminal recidivism? The authors use data on 148 youth who participated in SNPD and were under the jurisdiction of the juvenile court in one of six Texas counties. These youth also were diagnosed with mental disorders and had to meet two additional program eligibility criteria. Program participants were first compared to individuals on the SNPD waitlist and then to a comparison group constructed using propensity score matching. Because of low arrest counts, the authors use a negative binomial regression model to look at the effect of SNPD on total arrests and a Cox regression or duration model to look at the effects of SNPD on re-arrest hazards. The second JPAM article in this section is by Devaney and Fraker (1986). The authors use data from two waves of the Puerto Rico Food Consumption Survey to determine if the Nutrition Assistance Program (NAP), which cashed out food stamps, had any effects on food expenditures and diet quality relative to individuals receiving traditional food stamps. The authors employ traditional econometric methods as well as instrumental variable techniques and simulation models.

**Interrupted Time Series with Comparison Groups**

Author David Figlio (1995) examines the impacts of increases in the minimum drinking age on automobile crashes in Wisconsin using 18 years of monthly crash data. With such a long time series, the author can use Box-Jenkins methods to discriminate between long-term trends versus the effects of any declines in crashes that may occur because of increases in the drinking age. Wisconsin is unique in that it raised the drinking age twice over the relevant time period and had neighboring states that established and changed their drinking age at different times. An interesting secondary analysis in this study is an examination of the effects of border hopping by Wisconsin youths after accidents.

In another transportation related article, authors John Graham and Steven Garber (1984), examine the impact of the National Traffic and Motor Vehicle Safety Act of 1966 on death rates for vehicle occupants, pedestrians, and motorcyclists using data between 1947 and 1980. The authors find dramatic lifesaving effects of the Safety Act but cannot ascertain whether or not the new automobile safety features induced more reckless behaviors on the part of drivers. The regression approach used by these authors provides a markedly different analytical strategy from the Box-Jenkins model used by David Figlio.

**Posttests Only with Comparison Groups**

Two evaluations of the effectiveness of the Women, Infants, and Children (WIC) program and the ensuing debate that these articles engendered make them fascinating examples of the posttest-only design. By their very nature, because these studies examine birth outcomes, there is no obvious pretest (for example, birth weight before WIC).
Both studies examine the impacts of WIC, a program that provides food and nutritional advice to pregnant women and postpartum to women, infants, and children who are low income and nutritionally at risk. In the first study, Bitler and Currie (2005) use data on WIC participants from 19 states to examine the effects of WIC on a variety of outcomes, including gestation, birth weight, nights in the hospital (mother, child), prematurity, and whether or not the mother breast feeds. Joyce, Gibson, and Colman (2005) challenge Bitler and Currie's findings that essentially support the view that WIC works. They argue that these authors failed to account for the clinical literature that does not support a relationship between birth weight and nutrition, and they propose alternative outcome measures: birth weight adjusted for gestational age; a dichotomous variable equal to one if the infant falls below the 10th percentile for gestational age; and low birth weight for infants that are term. This study uses data from New York City birth certificates and analyzes both singleton and twin births. Joyce, Gibson, and Colman fail to find strong evidence that WIC works.

Because Joyce, Gibson, and Colman were at odds with the original article, I invited Bitler and Currie to comment on the later, contradictory publication. I also invited Ludwig and Miller (2005) to comment on the debate. Ludwig (then at Georgetown, now at the University of Chicago) comments as a policy analyst; Miller (Harvard School of Public Health) is a physician and public health researcher. This is an engaging and important debate that clearly demonstrates the sensitivity of program evaluation outcomes to the assumptions, measures, and estimation procedures used—even in sophisticated and nuanced studies such as those found in this section of this book.

Regression Discontinuity Design

In “An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs,” Wong et al. (2008) estimate a regression discontinuity model with the discontinuity (cutoff birthdates for pre-kindergarten) partitioning prospective students into treatment and comparison groups that are likely to be very similar, particularly for youths with birthdates close to the cutoff. The authors employ data from Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia children to determine if pre-kindergarten has a measurable, positive influence on students' reading and math skills. It should be noted that the regression discontinuity design is considered a particularly strong quasi-experimental design, albeit one that appears infrequently in JPAM (due largely to a paucity of submissions employing this technique).

META-ANALYSIS

A meta-analysis of 49 studies of environmental inequity conducted by Ringquist (2005) is included in this section. This author calculates effect sizes and estimates meta-analytic regression (MAR) models. He concludes that virtually all potential sources of environmental risk are disproportionately concentrated in areas with large percentages of racial and ethnic minorities, but that income is a far less powerful predictor of these potential risks.

Another meta-analysis that focuses on the value of a statistical life was also recently published in JPAM. There, authors analyze 33 studies and conclude that the value of a statistical life is approximately $1.5 to $2.5 million—although individual articles generate estimates ranging from $100,000 to $425 million. Because this article was included in the first JPAM Classic volume on cost-benefit analysis,
it is not included here. For instructors interested in using this article, they are referred to:


Also, the article “Do Experimental and Nonexperimental Evaluations Give Different Answers about the Effectiveness of Government-Funded Training Programs?” by David H. Greenberg, Charles Michalopoulous, and Philip K. Robins is a meta-analysis. This article is found in the section of this volume on social experiments versus quasi-experiments. In addition, the article “Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments” by Howard S. Bloom, Carolyn J. Hill, and James A. Riccio is a meta-analysis. This article is found in the next section on implementation, performance management, and program impacts.

IMPLEMENTATION, PERFORMANCE MANGEMENT, AND PROGRAM IMPACTS

Two studies are included in this section. The first attempts to look inside the “black box” to ascertain what aspects of programs lead to greater success. The article “Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments” by Bloom, Hill, and Riccio (2003), addresses the question of how implementation influences the effectiveness of mandatory welfare-to-work programs. The authors examine 59 local programs and 69,399 individual-level observations. They find that, ceteris paribus, earnings are improved by having participated in a welfare-to-work program that emphasizes quick employment, personalized client attention, reasonable staff caseloads, and limited emphasis on basic education.

The second study, by Barnow (2000), addresses a different but related question. Have attempts to quantify performance in the aftermath of the Government Performance and Results Act (GPRA) of 1993 resulted in performance measures that are related to programmatic impacts? Performance management is one way to hold programs accountable and sometimes reward high performers. In a sense, this quantification of performance is another lens through which we can examine process or implementation issues. This study uses experimental impact findings from 16 sites operating under the Job Training Partnership Act and looks at the extent to which higher “performance scores” correspond with more positive program impacts. Unfortunately, the author finds that while the relationship between performance scores and program impacts is usually positive, it is statistically weak—raising questions on the ability of the performance measures used in this study to capture relevant aspects of program implementation that impact individual outcomes. A more recent JPAM study on high performance bonuses and program outcomes reached similar conclusions. In fact, Heinrich (2007) concluded that high performance bonuses are more likely to promote falsification of performance and other strategic behaviors than to recognize and motivate outstanding performance.

ETHICS AND HUMAN SUBJECTS

Another issue in evaluation research receiving press concerns the extent to which participants in experiments and other evaluations are treated ethically—particularly as some “treatments” can be harmful—despite full compliance with requirements established by Institutional Review Boards (IRB) in universities and other firms (Curry, 2001; Brody, 2002). These IRBs have established detailed guidelines for the conduct of research designed to protect human subjects from unethical or negligent
behavior that could harm study participants and subject the research institution to legal action. They also test principal investigators on their knowledge of the guidelines to make sure that research analysts understand the rules and protocols relating to human subjects. Some argue that IRBs are performing an important function and are not imposing unreasonable restrictions on researchers, although they may be slow and overly cautious. Others debate what ethical standards should be used in evaluation research and if there should be formal mechanisms to ensure that evaluations are conducted ethically.

In this JPAM Classics volume, the discussion of ethics and human subjects taken from the Professional Practice section of JPAM. Jan Blustein (2005a) is physician and social scientist, began her work on this subject by looking at an application of the 1979 Belmont Report on medical ethics to social program evaluation. This article provides an historical overview of federal regulations that govern research ethics and documents how federal evaluation research has been specifically exempted from those regulations. The author poses but does not answer the following three questions:

- Are there good reasons to hold federal social program evaluations to different standards than those that apply to other research?
- If so, what ethical standards should be used to assess such evaluations?
- Should a formal mechanism be developed to ensure that federal social program evaluations are conducted ethically?

Burt Barnow, Howard Ralston, and Peter Schochet individually respond to Blustein article, after which Blustein responds to their commentaries. These readings are useful to stimulate and provide real substance to classroom discussions of evaluation ethics and human subjects.

THE USE OF PROGRAM EVALUATIONS BY POLICY MAKERS

In their article, "The Dissemination and Utilization of Welfare-to-Work Experiments in State Policymaking," authors David Greenberg, Marvin Mandell, and Matthew Onstott (2000) report the results of a survey of state-level administrators to assess the influence of three state welfare reform experiments on public administrators and policymakers. The programs considered were California's GAIN, New York's CAP, and Florida's Project Independence. While the experiments' impact assessments did not produce dramatic changes in policymaking, the authors found that the studies produced valuable information related to how the programs operated in the field, their political viability, and the states' ability to secure welfare waivers. This study highlights the importance of conducting implementation studies for public managers and policymakers.

CONCLUSION

The field of social experimentation and program evaluation is dynamic and multifaceted, with econometric debates on methods, a new stream of information from evaluations, meta-analyses of groups of evaluation, implementation studies, and discourse on ethics. In the end, I hope that readers will find this volume useful.

ACKNOWLEDGMENTS

I would like to thank Douglas Besharov, Thomas Cook, Laura Langbein, and William Zume for their comments on an earlier draft of this JPAM Classics volume. Of course, all opinions and errors are those of the author!
REFERENCES


THE ROLE OF RANDOM ASSIGNMENT IN SOCIAL POLICY RESEARCH

Richard P. Nathan, Editor

INTRODUCTION

This "Point/Counterpoint" exchange is on "The Role of Random Assignment in Social Policy Research," authored by Robinson G. Hollister, Jr., professor of economics, Swarthmore College, and Richard P. Nathan, co-director of the Rockefeller Institute of Government, who (necessary disclosure) serves as editor of this feature. The opening statements are published here. Each author's opening statement treats the following questions:

1. What is the proper role for random assignment social experiments? How generally and widely should they be used?
2. Are there alternatives to random assignment that are nearly as good? If so, what kinds of guidelines should be applied in deciding when to use random assignment and when to use alternatives?
3. What in your view are the best and worst experiences in applying random assignment to social policy research?
4. Looking ahead, what are fertile and, in your view, priority areas for social experiments and/or quasi-social experiments?
OPENING STATEMENT

Robinson G. Hollister, Jr.

1. What is the proper role for random assignment social experiments? How generally and widely should they be used?

Whenever it is desired to estimate the effect, or impact, of a social policy or institution on specific outcomes and one wishes to assert that the estimated impact is caused by the policy or institution, then a random assignment experimental design should be the method of first resort. The random assignment of units (persons, institutions, geographic areas, etc.) to the "treatment" or to a control group, if properly carried out with a large enough sample, assures that on average characteristics that might affect the outcome are the same (not statistically significantly different) in the two groups. This holds both for measured and unmeasured, or unmeasurable, characteristics. Further, as contextual conditions that might affect the outcome change over time (for example, a weakening of the labor market), both groups are subject to those changing contextual conditions. Therefore, the control group provides a reliable estimate of what would have happened to the members of the treatment group had they not been subjected to (or offered the opportunity to receive) the treatment; that is, the control group provides reliable counterfactual measures of outcome variables.

It is not always possible to mount a random assignment experiment. The impediments to doing so are varied, too varied to be listed here. One fundamental impediment arises where the treatment to be tested is already fully available to the population of interest and there is no legal or ethical way to withhold the treatment from potential control group members; excess demand for, or inadequate supply of, the treatment is a fundamental condition for applying random assignment.

"Random assignment can't be done here" is a statement researchers often hear at the outset. But strenuous efforts should be made to test arguments that perceived impediments are insurmountable. There are many ways to adapt the fundamental random assignment design to different contextual settings, and too often researchers and policymakers easily accept assertions of impossibility and slide back to other flawed methods of evaluation (discussed below).

The random assignment design, in general, provides "internal validity" of the impact estimates, that is, the causal claim for the impact estimates is internal to the study subjects. "External validity," the ability to generalize the estimated impacts to a broader population, is harder to assure rigorously. A careful delineation of the population from which the subjects are drawn, and that they are drawn randomly from that population, are prime requirements for a claim of external validity. As far as I know, few major experiments could rigorously meet this standard. But this is also the case for most nonexperimental studies—in addition to the limitation of not being able to substantiate internal validity.

2. Are there alternatives to random assignment that are nearly as good? If so what kinds of guidelines should be applied in deciding when to use random assignment and when to use alternatives?

The short answer to the first of these questions is: No, not as far anyone has been able to establish thus far.
There have been claims for particular methods being "nearly as good," but with perhaps one exception, discussed below, none of these claims has held up to close scrutiny. There is an important literature that has built up over the last two decades that tests nonexperimental methods of estimating impacts against actual experiments to see how close the nonexperimental methods can get to the impact estimates derived from the actual experiment. One should recognize that, at least in theory, how "close" an estimate from the nonexperimental method comes to an experimentally derived estimate is likely to vary according to the type of treatment being addressed and the types of populations or places to which it is applied. However, the conclusion has been pretty much the same across these areas: It is hard to establish a second best nonexperimental method.

The proposed nonexperimental methods have ranged from straightforward regression analyses to complex matching schemes. In the straightforward multiple regression analysis using existing data sets, some variable is identified as reflecting the "treatment" of interest (for example, a variation across states in the level of welfare benefits), and it is used as the independent variable, with the dependent variable being the outcome measure of interest (for example, the level of earnings). All other measures of characteristics are entered in the regression as control variables, hopefully controlling for differences between those receiving and those not receiving the treatment. The most recently touted and investigated complex method is called "propensity score matching." I will not attempt to describe it here, but it has been thoroughly tested in the cited literature.

The base problem seems to be that the nonexperimental methods do not deal adequately with unmeasured or unmeasurable variables (for example, motivation), which can affect the outcome and are correlated with the treatment, positively or negatively. When I said to the creator of one of the most complex nonexperimental methods, "But this method doesn't deal with the effects of unmeasured variables," he said, "Well, I guess those effects are in the hands of God." A random assignment procedure gets around this problem because such unmeasured variables will on average be the same in the treatment and control group and therefore not confounded with treatment status.

One nonexperimental evaluation method that does seem to hold promise (the exception noted above) is Regression Discontinuity Design. This design has been recently used in several major studies, and there are several replication studies against experiments that show that it gets closely similar impacts, although it requires about three times the sample size for equivalent power.

A major advantage of a random assignment design evaluation is that it is easy for nontechnical people to understand the results: One compares the mean outcome for the treatment group to the mean outcome for the comparison group; there are no complex technical assumptions to be understood.

There is another point to be made with respect to the choice between random assignment experiments and alternatives to random assignment. It is often said that experiments are much more expensive than nonexperimental methods. Even researchers who carry out experiments often accept this argument with little protest.

---

1 Glazerman, Levy, and Myers (2003); a recent overview of such studies as applied to employment outcomes. Other important studies in this regard are LaLonde (1986); Fraker and Maynard (1987); Friedlander and Robins (1995); Heckman, Ichimura, and Todd (1997); Michalopoulos, Bloom, and Hill (2004); Agodini and Dynarsky (2004); Smith and Todd (2005); and Wilde and Hollister (2007).


3 I do not have space here to discuss instrumental variables (IV) designs nor "natural experiments." For an excellent and accessible discussion of strengths and weaknesses of these approaches, see Angrist and Krueger (2001). For a seminal discussion of the use of the IV approach as a complement to random assignment studies see Angrist, Imbens, and Rubin (1996).
The problem is that this simply is not true. There is nothing about using random assignment that makes it more expensive than alternatives. Think about it: Random assignment requires no more than the equivalent of a coin flip. The basic determinant of the costs of an evaluation is the cost of the data to be gathered and, in particular, the costs of original data collection—data generated in the course of the evaluation. On first consideration, it might seem that if one can draw a comparison sample from an existing data set in order to match with a treatment group and estimate impacts, it would be cheaper. But the problem, as shown in the just cited literature, is that it is critical that the measures of the variables in the comparison group and the treatment group, especially the outcome variables, be as close as possible; for example, educational achievement measured with the same test instrument or earnings defined and measured in the same way.

3. What, in your view, are the best and worst experiences in applying random assignment to social policy research?

Best Experiences

**Negative Income Tax Experiments (1968–78).** I refer to a set of four experiments testing the effects of a proposed income transfer program commonly referred to as the Negative Income Tax. These were, to my knowledge, the first use of a random assignment design on a large scale to attempt to estimate the effects and related costs of a proposed new policy in the welfare/income maintenance area. As researchers at the U.S. Office of Economic Opportunity attempted to estimate the costs to the federal budget for this income support, a central question was how the receipt of such income transfers would affect the work effort (labor supply). Using the then-available data and econometric techniques, the researchers got a ridiculously wide range of estimates of the effects. It was decided that the only way to get reliable estimates was to do a random assignment experiment.

These experiments really set the stage for the whole series of social policy experiments that have followed. The researchers developed important basic elements required for good experiments—sample design and selection, clear details of treatment, original data collection including strong follow-up, and methods of analysis that subsequent experiments built on.

The experiments did have an impact on federal policy discussions. Premature results from the New Jersey experiment were cited in the debates over Nixon’s Family Assistance Plan, a form of negative income tax (which most people don’t realize passed in the U.S. House twice, only to be defeated both times in the Senate). The results were widely cited in the debate over President Carter’s welfare reform (discussed further below). It can be argued, I believe, that these experiments really drew the attention of the research community, particularly economists, to the problems of developing impact estimates that could establish a causal link from the policy to the outcome. Over the ensuing decades, development of methods that explicitly attempt to address these problems have slowly played a larger role in quantitative work, particularly, but not exclusively, in the social domain.

**National Health Insurance (1972–82).** When this experiment was initiated, many thought we were on the brink of a National Health Insurance program. A central issue was, Should any national health insurance provide fee-free health services or

---

4 One slight possible excess cost element in an experiment is the cost of finding subjects (individuals, schools, etc.) who agree to participate in the experiment and abide by the random assignment.

5 I note, to warn against my personal bias, that I was deeply involved in the design, implementation, and analysis of the first Negative Income Tax experiment (New Jersey-Pennsylvania), and I was the co-principal investigator for the evaluation of the National Supported Work Demonstration.

6 See Greenberg, Linksz, and Mandell (2003) for in-depth discussion of several of the experiments listed below.
should there be cost-sharing provisions (deductibles, coinsurance)? Members of the treatment group were assigned to one of 14 different fee-for-service plans, including a fee-free plan and an HMO plan. It has been argued that the results have played a role in the development of the managed care method of health services provision.7

Supported Work (1974–80). Supported Work provided subsidized direct employment for a period of up to 18 months for members drawn from four populations: unemployed ex-offenders, former drug addicts, women who were long-term welfare recipients, and young school dropouts. MDRC was created to manage the project and provide technical assistance. The result that surprised the designers was that the program was most effective in raising the post-program employment and earnings of women on welfare. It can be argued that it led to the subsequent experiments in welfare work programs that eventually culminated in the welfare reform of 1996.

Given limited space, hereafter I only briefly note other major experiments I think represent “best experiences.”

MDRC Welfare to Work (1985–2001). This includes a series of random assignment studies, mounted jointly with states and localities, that had major effects on the Family Support Act and the 1996 major national welfare reform. They tested the work-first vs. human capital approach to employment enhancement, showing that the work-first approach was more effective.

National JTPA (1986–93). The random assignment study of the ongoing national job training system. Its findings led to a major revamping of the U.S. Department of Labor’s approach to job training.

Project Star (1985–89). The single major random assignment study in the education era prior to 2000. The results have had a big influence on the class size debate. In addition, the study was the source for other rigorous analyses, for example, effects of race matching of students and teachers on achievement scores.8

Job Corps (1993–2003). The evaluation involved a true random sample of Job Corps applicants across the country and as such yields one of the few studies for which external validity can be claimed. The design was elegant and implementation excellent. When records data (social security) were analyzed, it resulted in substantial revised conclusions about the program’s impacts.

Abstinence Education (1997–2007). This random assignment test of abstinence-only sex education programs was executed in a highly politicized environment. The results, showing little effects in terms of sexual behavior; received immediate national attention and had an effect on policy.

Worst Experiences

Perry Preschool (1962–1987).9 The High/Scope Perry Preschool Project (hereafter Perry) was an early childhood program targeting low-income children. Children attended the preschool two-and-a-half hours a day five days a week, and there were weekly home visitsations with their parents. The evaluation was based on a random assignment design. Follow-up measurements were taken until the subjects were age 27.

For decades this has been the most widely cited evaluation, which asserted there are powerful, long-lasting effects of early childhood education on academic, earnings, and other social outcomes, most notably crime.

Although initial efforts were made to implement a standard random assignment, the researchers made several nonrandom adjustments to the assignment, for

---

7 See http://www.rand.org/pubs/research_briefs/RB9174/, The Health Insurance Experiment, for an overview and some speculation as to applicability of the results to present situations.

8 A timely review of research derived from Project Star can be found in Whitmore Schanzensbach (2007).

9 For an excellent review of the features of, and problems with, Perry Preschool see Besharov, Germanis, and Hingley (2007).
instance, moving siblings so that they would be together in the treatment or control
group, or moving all children of working mothers to the control group. In the end,
20 percent of the sample that was used in the subsequent multiyear follow-up eval-
uation were not randomly assigned. One cannot tell which direction these deviations
from pure random assignment may have biased impact estimates, in favor of or
against the program; for me, these failures of random assignment greatly undermine
one's ability to take estimates of the "impacts" as sound.

Further doubts about the reliability of the estimates arise from the fact that the esti-
rated impacts in given areas, for example, achievement test scores, vary sharply over
time (age of the child). For instance, the crime data suddenly show big differences in
favor of the program in the age 27 data. The estimated impacts on crime play a large
role in the overall high benefit-cost ratios that have been highly touted.

What is most disturbing to me is that the findings from this study have been the
most often cited evidence of the great benefits to individuals and society that flow
from early childhood programs.\textsuperscript{10} The program has not been replicated and rigor-
ously evaluated in other studies.

\textit{Seattle-Denver Marital Stability Analysis.} As part of the Seattle-Denver Income-
Maintenance Experiment (one of the Negative Income Tax experiments discussed
above), analysis was done to estimate the effects of receipt of the Negative Income
Tax payments on marital stability (either divorce or marriage) of those couples
receiving payments. The analysts concluded that payments increased marital insta-
bility. These results were surprising and at variance with those from the other nega-
tive income tax experiments (which showed no marital stability effects). Congress
was considering President Carter’s proposed welfare reform—which took, in part,
the form of a negative income tax—when testimony about this finding was pre-

t\textit{sent to them. According to many observers, the testimony led to the end of any
congressional interest in Carter’s reform proposal.\textsuperscript{11}}

The tragedy is that the research conclusions were wrong. But it was 10 years before
the corrections to the analysis were completed and published.\textsuperscript{12} Cain and Wis-
soke painstakingly reconstructed the data from the experiment, adjusted the estimating
model and procedures, and showed that in the long run there was no statistically
significant marital effect.

The damage was done to the welfare reform effort by those initial estimates, but
within the broader research community the belief persists that an increased mar-
tial stability effect from income transfer payments was shown to exist. Though
published and debated in a prestigious social science journal, the corrections have
not been noticed by the bulk of the research community. Sadly, when poor social
science analysis gets major publicity, it is virtually impossible to get those con-
clusions out of the general discourse and the correct conclusions substituted for
them.

\textit{The Quantum Opportunities Program (pilot) (1989–93).} The Quantum Oppor-

tunities Program (QOP) was a year-round, multiyear, comprehensive service program for
disadvantaged youth launched in five communities in 1989. Twenty-five disadvan-
taged students in each community were randomly selected to enter the program
beginning in ninth grade and continuing through four years of high school. QOP was
focused around education activities (tutoring, computer-assisted instruction) and
development activities (life skills, planning for the future including postsecondary
education and jobs). Community service was also stressed. Young people were pro-
vided with adult mentors who stuck with them over four years. Financial incentives

\textsuperscript{10} Further studies in other areas, for example, nonexperimental studies of after-school programs, have
used these results, particularly crime effects, to "show" that such programs yield huge benefits to society.
\textsuperscript{11} Senator Patrick Moynihan, who had been a supporter of Carter’s proposals and Nixon’s Family As-
sistance Plan (a negative income tax-type plan), was especially shocked.
\textsuperscript{12} See Cain and Wissoke (1990).
were provided, through stipends and bonuses, to participate in QOP activities.\textsuperscript{13} It was an intensive program that helped students for up to four years, and the costs per enrollee were very high, ranging across sites from $18,000 to $49,000.

A summary of the evaluation of the pilot was compiled,\textsuperscript{14} but that evaluation completely misrepresented the underlying research data.\textsuperscript{15} The random assignment was flawed in several of the sites. Only one of the five sites—one with correct random assignment—had statistically significant effects, but they were presented as though they applied to all the sites; further, effects estimated in the flawed sites were presented as occurring across all the sites. These results continue to be touted on several Web sites, and the Department of Labor for many years recommended the program model to those wanted to field “an effective youth program.”\textsuperscript{16}

4. Looking ahead, what are fertile and, in your view, priority areas for social experiments and/or quasi-social experiments?

First, the Institute for Educational Science, U.S. Department of Education, was created in 2002. It announced that the first priority for the research studies it funds—both through its grant programs and congressionally mandated evaluation—would be those that used random assignment design. At the time, the field of education had virtually no such studies (Project Star excepted). These efforts to fully utilize random assignment designs deserve the greatest support possible.

Second, in my view, the American workers are experiencing much greater insecurity due to changes in labor market institutions and norms. These are most readily recognized in terms of increasing lack of health insurance, weakening of the pension system, and related retirement concerns. Under this general heading, there are several topics that deserve priority studies using random assignment designs.

\textit{Financial education.} Given weakening of employer and union-based policies, workers are increasingly on their own to deal with management of the financial dimension of their security. Testing the effectiveness of alternative forms of financial education should have high priority.

\textit{Savings and pension vehicles and strategies.} Obtaining more rigorous information on how to encourage savings and better decision making about pension and participation are critical.

\textit{Wage subsidies.} Although there have been a couple of experiments\textsuperscript{17} partially testing wage subsidies as a means to increase employment and income of low-skilled persons, full-scale testing of these means of increasing employment of low-skilled persons needs to be done.

Third, considerable work is already underway on what I would call “behavior-based transfer payments.” The prime example of this is the Progressiva experiment in Mexico, in which transfer payments to poor households are contingent on continued school attendance of children, utilization of health clinics, and improved nutrition practices. This type of program is being tried in other Latin American countries. Mayor Bloomberg has proposed a version of it to be tried in New York City. It is important to determine through a random assignment evaluation whether the sorts of

\textsuperscript{13} The above descriptions are drawn from the Promising Practices Network: Quantum Opportunity Program (www.promising.practices.net) and Quantum Opportunities American Youth Policy Forum (http://wch.uhs.wisc.edu).

\textsuperscript{14} See Hahn, Leavit, and Aaron (1994).

\textsuperscript{15} Details of the misrepresentations are available on request.

\textsuperscript{16} In 1995–2001, the Department of Labor and the Ford Foundation funded a replication of a version of QOP in seven sites with a random assignment evaluation. The overall impacts were not statistically significant for the primary objective, increasing the likelihood of graduating from high school with a diploma or GED. See Schirm, Stuart, and McKie (2006).

\textsuperscript{17} New Hope Project, Milwaukee, Wisconsin and Canada’s Self-Sufficiency Project. For further discussion see Berlin (2007).
results found in Mexico would be to some degree replicated in the context of a developed country urban setting.

Fourth, following the tremendous growth in prison populations in the 1980s and 1990s, we now face a large wave of prisoners completing sentences and attempting to reenter the mainstream. We need to have a systematic set of experiments to test alternative reentry programs.

REFERENCES


OPENING STATEMENT

Richard P. Nathan

1. What is the proper role for random assignment social experiments? How generally and widely should they be used?

I am bearish on random assignment. That is not to say negative. I believe social experimentation with random assignment has a proper, but limited, role and that this role is not widely enough and well enough understood by the sponsors and practitioners of public policy research.

Supported Work Experiment

It was Robinson G. Hollister, Jr., who educated me on this subject when I participated in the founding of the Manpower Demonstration Research Corporation (MDRC) in 1973. Rob was the lead consultant on the design of the initial MDRC supported-work demonstration. The essential question for the supported-work demonstration was how to aid four groups in need of social assistance. For one group, single mothers, MDRC found significant positive impacts of intensive federal supported-work programs. This was a well-managed, closely monitored five-year demonstration research project using random assignment, conducted between January 1974 and 1979 at a cost of $82.4 million. Such investments of money and social science expertise in social experimentation are proper—but in my view only under the following conditions: (1) when leaders in government are undecided about what to do and are interested in finding out; (2) when they are willing to spend a relatively large sum of money to do so; (3) when they test policies and interventions that reasonably can be expected to have a significant impact; (4) when they are willing to wait for the results; and (5) when ethical issues raised by random assignment are dealt with in an acceptable way.

Abraham Kaplan, a philosopher of social science, used the metaphor of giving a child a hammer and having the child soon find that everything needs pounding. I regard social experimentation as an important tool, but not for every job or object in the house. In the MDRC case, Robert Lampman, an influential MDRC board member, taught us that what I think is an especially valuable lesson. He urged us to beware of the null hypothesis—that is, testing treatments in situations in which there is not sufficient reason to believe that they are large enough and will be sustained long enough to be expected to change outcomes in something as complex as a human life (that is, condition number 3 above). Lampman said, “When persons with severe employment handicaps and disabilities are singled out for remediation, positive and lasting effects are not likely. In the case of supported work, the odds in favor of the null hypothesis were even greater... since the four groups chosen were from among those least likely to succeed in the labor market.”

---

18 See The Board of Directors, Manpower Demonstration Research Corporation (1980). This research was supported primarily by the U.S. Department of Labor; with initial seed money from the Ford Foundation.
Aaron Critique

Lampman’s comment fits with Henry J. Aaron’s critique that formal social science in too many instances has undermined the case for social programs. Assuming that most researchers believe not only in the utility of applied social science, but also in the value of social programs, this is a gloomy situation. Said Aaron: “The role that research and experimentation played in the demise of the simple faiths of the early 1960s was not accidental. The process by which R&E (research and experimentation) is created corrodes the kind of simple faiths on which political movements are built.”

Even when the listed conditions apply, there are situations in which formal social experimentation is not appropriate. In the case of the negative income tax (NIT) experiments, researchers were testing a big-picture reform, which, if it had been enacted, would have changed values and expectations in the society as a whole. In my view, we cannot test a universal reform like an NIT in a vacuum. Its enactment could have been expected to change work-seeking behavior, as Daniel Patrick Moynihan told his friends at the time he hoped it would, although he didn’t say so in public. The same is true of the global-signaling effect of universal health reform plans. Adopting fundamental change on big-ticket components of the nation’s social contract is likely to contaminate the setting for classical demonstration research.

Balance Needed

Looking broadly at how social science can best be applied, my concern is about balance. Random assignment is so often highly favored in the academy as the purest (often referred to as the “gold-standard”) application of social science in public affairs that this has had a distortion effect. Its focus on what works has been at the expense of other critical areas for public policy research, for example, studies of social and economic problems (what needs to be done) and implementation studies of the capacity of existing, ongoing public programs (what governments actually do).

2. Are there alternatives to random assignment that are nearly as good? If so, what kinds of guidelines should be applied in deciding when to use random assignment and when to use alternatives?

This question about alternatives to random assignment is frequently treated in the academic literature in relation to nonexperimental techniques. I want to stretch the canvas. There are two major and competing paradigms that policymakers use to assess what works. One is a Weberian approach highlighting rigorous and presumably replicable social experiments and quasi-experiments. The second alternative is the performance management movement, which is discussed next.

There is a litany of good-sounding alphabet-soup systems that governments adopt and invest in to measure program performance and rate, and rank government agencies and ongoing programs. PPBS was Lyndon Johnson’s Planning-Programming-Budgeting System; MBO was Nixon’s successor approach for Management by Objectives; ZBB was Carter’s more radical system for Zero-Based Budgeting; NPR stood for Clinton’s National Performance Review. A national government requirement for a government-wide system for performance management was enacted into law in 1993 in the Government Performance and Results Act (GPRA). The latest such system is the PART system under George W. Bush, which has been used extensively to justify program eliminations and budget cuts.

---

22 PART stands for “Program Assessment Rating Tool,” a system to compile and publish information and ratings on the effectiveness of all federal programs.