Learning What Works:
Evaluating Complex Social Interventions

Report on the Symposium

held on
October 22, 1997
The Brookings Institution
Washington, D.C.

The Brookings Institution
Governmental Studies Program

And

Harvard University
Project on Effective Interventions

March 1998

Foreword

A day's thoughtful discussion among a group that came at the topic of "Evaluating Complex Social Interventions" from many varied perspectives, produced a number of shared observations and insights.

Perhaps the most notable was the idea that while efforts to provide information about complex social interventions rarely produce certainty about "what works," these efforts can and should be constructed to provide useful clues to what *may* work and what is promising. Both social scientists and practitioners involved with comprehensive community initiatives and other complex interventions may be well advised to avoid claiming "We know what works!" Rather, they should aspire to gathering and analyzing the information that would enable there to say, "We now have strong support for a number of informed hypotheses about what may work, which cumulatively, over time, can produce sturdy knowledge about what does work."

Several participants pointed out in various contexts that where outcomes improve substantially in areas that the public cares about (such as a significant increase in the number of inner city children leaving school prepared for employment), issues of evaluation methodology tend to move into the background.

The group also seemed to agree on the need for greater recognition of the importance of the values, norms, and rules of behavior that has generally fallen outside the realm of economic models of evaluation.

Few challenged the contention, voiced in various forms throughout the day, that political considerations often trump rational analysis of effectiveness in making policy judgments. Despite this caveat, and while many participants expressed reservations about the current state of evaluation and how it is used in assessing complex interventions, there was strong support throughout the symposium for greater and more systematic investment in utilizing and integrating a variety of methods to compile useful, rigorous information about the operation and effectiveness of complex interventions. But such investment would only be warranted, several participants warned, when (1) there is clarity about expected outcomes, and the theories connecting interventions, interim markers, and outcomes, and (2) the interventions are functioning at a scale and level of intensity that make it reasonable to believe that they may be successful.

There was also considerable sentiment in support of the idea that greater investment in bold interventions themselves was warranted, because "you have to try an awful lot of things to find out what does work," or even what might work.

We trust that the symposium will stimulate participants and other colleagues to pursue further the provocative ideas that the discussion generated. To this end, we hope that the following summary of the proceedings -- prepared by Kathleen Sylvester of the Social Policy Action Network, with final editing by the two of us – will prove useful.

As with all Brookings publications, the opinions expressed in this report are those of the authors and should not be attributed to the trustees, officers, or other staff members of the Institution.

Thomas E. Mann
Director, Governmental Studies
The Brookings Institution

Lisbeth B. Schorr
Director, Harvard University Project on Effective Interventions Co-Chair, Aspen Institute Roundtable on Comprehensive Community Initiatives

Contents

*(page numbering for this PDF version differs from previous printed copies)

Report on the Symposium

At a meeting convened by the Brookings Institution's Governmental Studies Program and the Harvard University Project on Effective Interventions, a small group of invited participants (see Appendix A) met on October 22, 1997, under the chairmanship of Thomas E, Mann, director of the Brookings Governmental Studies Program, to consider haw evaluators can better provide credible and generalizable information about the effects of increasingly complex social interventions.

The need for reliable information is urgent because the problems that complex social interventions are trying to address -- from urban decay to family dysfunction -- are urgent. The need is urgent because both public and private funders and front-line people implementing complex interventions are desperate for feedback that will guide their efforts.

At the same time, an increasingly skeptical public is insisting on proof of results. That public includes not only the taxpayers and foundations that pay for these interventions, but also the community residents involved in trying to make the efforts succeed. They want to know that what they are doing will produce the promised results. The need for credible and objective evaluation has been heightened by the tendency of governments, particularly the federal government, to become more and more immobilized with partisan bickering and special interest politics.

This was the context outlined by Lisbeth B. Schorr, who drew on her new book, *Common Purpose, Strengthening Families and Neighborhoods to Rebuild America*, to offer the hypothesis that some of the most promising interventions are least likely to be understood with the methods most commonly used -- and considered most credible -- for evaluating them. Schorr suggested that traditional evaluation techniques may not be well suited for measuring the success of interventions with the most promising characteristics. She also warned that the limitations on [?] prevailing evaluation methods can inhibit promising experimentation when sponsors worry about designing programs to make them "evaluable."

Schorr listed the characteristics of promising interventions that have proven difficult to evaluate with traditional approaches. They intervene in more than a single facet of people's lives and change more than one thing at a time. They are designed to impact not only individuals, but also neighborhoods, institutions, and systems. They differ from one community or neighborhood institution to another, reflecting

---

[?] The Symposium was planned and sponsored jointly by the Brookings Governmental Studies Program (Thomas E. Mann, Director) and the Harvard University Project on Effective Interventions (Lisbeth B. Schorr, Director), and funded by the Annie E. Casey, Joyce, Swing M. Kauffman, and John D. and Catherine T. MacArthur Foundations.

differences in local needs and strengths and the adaptations or inventions that make for local ownership. They encourage flexibility on the front lines. And they evolve over time in response to experience and changes in local needs.

Citing reducing teenage pregnancy and rebuilding impoverished neighborhoods as examples of tasks requiring complex interventions, Schorr suggested that it "is close, to impossible to learn and generalize about the impact of those kinds of complex interactive interventions from assessments that are considered credible by most policymakers." The unfortunate result, she said, is a lack of knowledge needed for program design and policy decisions. She referred to former Office of Management and Budget director Richard Darman's observation that "the great Washington scandal is that the federal government embarks on domestic programs without reliable, empirical evidence that they would work." In a 1996 article in the *New York Times Magazine,* he called on President Clinton to "end the era of policy corruption by initiating a set of bold research trials focused on major policy initiatives to guide social policy in the 21st century."

To accept Darman's challenge, she said, social scientists would have to be "willing to include within the charmed circle of credible knowledge the understanding that requires intelligent judgment to make up for some absence of certainty." She suggested that the social scientists who have taught policymakers to dismiss information that may be relevant, timely, and informative, but lacks certainty, have contributed to the widespread sense that nothing is known about what works -- because the certainty we demand is not attainable. Thoughtful observers with access to a wide array of data about what happened and what might have happened under different circumstances can combine that information with an understanding of similar interventions and build a strong and useful knowledge base, said Schorr.

While conceding that such a body of information would lack certainty about causation, Schorr predicted that it "would be more likely to lead to effective action on urgent social problems than conclusions based on a narrower definition of what is credible and scientific."

In responding, Henry Aaron of the Brookings Institution raised several issues. First, he noted that human behavior shifts in unpredictable ways because of the values, norms, and behavioral rules that people change as a result of their social interactions with peers. He cited the resurgence of interest in using religious institutions as instruments for change because of the ability of churches to influence the commitments, values, and norms that are at the root of changing human behavior.

Some behavior, said Aaron, is not incorporated in traditional economic models. For example, the sources of the norms and habits that keep citizens from cheating on their income taxes and lead them to give to charity are more likely to be social interactions, families, and church, not investments.

Aaron also pointed to the emergent field of computer-based modeling and modeling of artificial life. These, he said, demonstrate how behaviors emerge from

group interaction. A possible example comes from the history of Social Security retirees. Their behavior has responded very slowly to dramatic changes in the rules and incentives the system creates for its beneficiaries. Social Security rules were changed in the 1960s to allow beneficiaries to retire at age 62 with reduced benefits, few people took advantage of the change. Now, nearly 70 percent of men retire before age 65.

This suggests that changing the rules will affect the behavior of some people. Those changes in behavior will affect others, and the process can go on for an extended period of time. The problem for evaluators, said Aaron, is that those who look at cross-sectional behavior do not get the pattern of the way things play out over time.

How Values, Optimism, and (Mostly) Politics Compromise Honest Evaluation

Even if analysis could take into account the complexities of the intervention and its assumptions, asked Aaron, what happens when the evaluation shows that the intervention did not meet expectations? What is the policy implication when an experiment fails? One can, of course, stop the experiment or program -- that is, stop spending money on something that does not work. But that does not solve the problem the experiment or program was designed to address. "Knowing that something didn't work," noted Larry Orr of Abt Associates, "doesn't tell you what would work."

If the problem is too important to abandon, policymakers are likely to try something else -- whether or not it is known to work. Thus policy decisions may not be based on evaluation, said Aaron, but on social and political judgments about the seriousness of the situation that flows from the decisionmakers' prior judgments and commitments. Mann suggested that it is possible to rationalize a failed result by hypothesizing that an adjustment in the intervention would have made it work, a phenomenon he labeled a "positivity bias." Orr suggested that such modified interventions should be rigorously evaluated before being adopted on a large scale.

William A. Galston of the University of Maryland raised the issue of another kind of bias: a political bias. Pointing out the political resistance to his own proposal for limited "field testing" of school vouchers in ten urban areas, Galston told the group: "There are a lot of people who resist that strategy, not because they're afraid this particular idea is going to fail, but because they're afraid it's going to succeed. And a lot of that is driven by politics and ideology."

While some resist evidence of success of ideas they oppose ideologically, politicians cannot resist, "magic bullets." Kent Weaver of Brookings spoke of politicians' rush to turn small effects into magic bullets. He cited the success of welfare-to-work programs for some welfare recipients as the reason for policymakers' nearly singular focus on the "work-first" approach. Weaver also reminded the group that politicians flocked to interventions such as family caps and time limited welfare

benefits, in part because there were no evaluations showing that these interventions did not work.

Mark Schmitt of the Open Society Institute noted that while there are many examples of Aaron's cases where negative evaluations result in giving up on a project, and of Weaver's cases where small effects are magnified into magic bullets and used to justify large policies, there are more cases in which very important and useful evaluations are completely ignored in the policy process.

Schmitt, who had served as policy director for former Senator Bill Bradley of New Jersey, suggested that this is particularly true on Capitol Hill as opposed to the Executive Branch. But it is important to pay attention to Capitol Hill because it is in Congress where things get scaled up and where fads take off.

Commenting on Bill Galston's observations about school vouchers, Schmitt suggested that no one's mind is going to be changed about school vouchers, regardless of what evaluations suggest. Said Schmitt: There are a lot of things that policymakers do because they regard those decisions as representing basic values. Alluding to Weaver's remarks, Schmitt said that a good example is New Jersey's cap on benefits for families on welfare, which denies welfare mothers higher payments for additional children born after the mothers are on welfare. Evaluation results, he said, really did not matter.

Darman suggested another political consideration. Whatever the informed answer turns out to be, he said, putting a fair price tag on the costs of the research is likely to be very much more expensive than the political community is willing to tolerate at the moment. Contrasting the government's spending on social policy research to its spending on medical research, Darman asked why a failing education system or a failing health system should be regarded as any less important in allocating research funds than cancer or AIDS.

Brookings' economist Gary Burtless pointed out that good evaluation can be the friend of good program design, not its enemy. He suggested that programs that pass a convincing and reliable evaluation are more likely to attract funding from both private and public sources than those for which no tangible evidence of effectiveness exists. This is especially true of programs that have weak political constituencies. Policymakers depend on evaluation results for the support of programs for which there is a weak demand. And he suggested that because the goal of many comprehensive community initiatives is to affect populations beyond the target groups that receive services or interventions, it is important to be able to measure the indirect effects or "positive externalities" on other people in the same community.

And finally, he cited the example of comprehensive employment training programs. Evaluations of those programs yielded the information that one extremely inexpensive component of the program -- jobsearch assistance -- produced better results than the more expensive components. The information was valuable because when

budgets to help disadvantaged people were scaled back in the 1980s, job-search funds were actually increased.

He noted that another reason to distinguish the separate effects of programs with multiple components is that if some of the components are not politically "saleable," there may be reliable evidence that others are effective.

Orr agreed that policymakers are far less tolerant of broad experiments in social reform. He noted that, if most social programs do not work, "then you have to try an awful lot of things to find out what does work. And we don't do that. We try one thing at a time." And when we have tried multiple interventions simultaneously – as in the case of the numerous youth employment demonstrations of the late 1970s – there has been little solid evaluation.

Evaluation for What?

Galston suggested that the lack of solid evaluations extends even to very large, very well established, longstanding public programs such as Head Start. Moving to an even more provocative issue, Galston used the example of the District of Columbia's complex intervention to improve its failing public schools, to pose the question of "What is it we're really interested in finding out?" He suggested that if by the year 2001, the city's schools were safer, student achievement had improved, and rates of graduates' admission to post-secondary education and training had increased no one would be curious to know the evaluators' opinions on why that had happened. "For political purposes, for public purposes, what the American people want is a package that moves the needle along the direction that they care about – and along the dimensions they care about." And if that happens, he said, "that's enough."

Orr, articulating the concerns of evaluators with political reasoning, raised two issues: causality and budgets. First, Galston's "counterfactual" – his extrapolation of the 1990 to 1997 performance of the D.C. public schools – should be examined to rule out possible effects of demographic changes or population shifts between the city and the surrounding suburb. Next, the interventions should be examined to determine if one particular intervention or combination in the package caused the change. If an intervention is complex and expensive, he said, it might be useful to determine what elements of the intervention can be eliminated to save money.

Isabel Sawhill of Brookings offered three potential purposes for evaluation: using evaluations as evidence to change public opinion; making decisions about whether to invest more money or less money in a given approach; and providing "feedback" for mid-course corrections. Anne Kubisch of the Aspen Institute's Roundtable on Comprehensive Community Initiatives added a fourth. She agreed with William Galston's contention that if the changes in outcomes for District of Columbia schools improved dramatically, no one would insist on knowing which components of the intervention caused the change. "It's when the magnitude of change is much smaller that

we've tended to fall back on the sophisticated, analytical tools that the scientists have given us."

As an example of even "inelegant" evaluation results that can inform policy making, Schorr cited a recent evaluation of Savannah's Youth Futures Authority. Outcomes of this complex, community-wide initiative, including repeat teenage pregnancy, healthy births, and family formation, were all going in the right direction. But school outcomes were not. The data, which compared the target area in Savannah with the surrounding county, were relatively simple, said Schorr. But knowing that the intervention components did not include reforms aimed at improving classroom teaching and learning, would it not be possible to hypothesize from this data that if the leaders of the effort wanted to change school outcomes, they might have to include reforms aimed specifically on changes in the classroom?

Again illustrating the differences in perspectives among the group, Swarthmore College's Robinson Hollister Jr. challenged Schorr to explain how generalized knowledge from the raw data from Savannah could provide insights to another community. Schorr responded by suggesting that the people in Savannah can say we did "these five things, and they resulted in changing these outcomes, but not those." Schorr contended that that would be an interesting conclusion -- not only for Savannah but for other communities.

Is There Any Consensus?

Despite these differences in perspective, a number of general points of agreement began to emerge from the discussion. As articulated by Darman, these points included:

- That the problem is not most importantly methodological, though there are methodological problems;

- That efforts at social change would benefit from more bold initiatives, and should include larger variations in the initiatives intended to address the same particular problem; and

- That there is a need for greater resources to fund both these bolder and more varied initiatives and the varied methodologies that should be applied to their evaluation.

Darman also urged the group to use some slightly humbler language than "learning what works." Perhaps we should be talking about "learning what may work, learning what some people think may work. That's a stage you have to be at on the way to learning what works." He suggested that jumping prematurely to say "we know it works" stirs up distractive controversy about methodology. "If you're slightly more humble and you say that now we really have support for an interesting hypothesis, then it's informed judgment. And over a much more extended period of time, the society

increases its intellectual capital and works its way towards knowledge about what does work."

Darman added a plea for vouchers and for decentralization, because both create a framework in which there is more variation, more opportunities for evaluation in relation to set measures, and more opportunity to learn what may work over time.

The discussion turned to politics again, as Darman observed that liberals are generally more interested in evaluation than conservatives, because liberals are more interested in making government work. Conservatives who would like to kill certain programs do not care about improving them or knowing what works. If you believe that the government has no role in a particular area, there is no need to spend money to figure out how to make government programs work better in that area. He suggested that if those who wish to preserve the government role and improve programs were to add to their agenda some tolerance for vouchers and decentralization, there would be "a potential for a marriage there that could produce a lot of what you'd want in the way of institutional change."

Reports from the Field

Mann then asked Paul Hill, Robinson Hollister, and Peter Rossi to highlight what is currently going on with interventions and evaluations that is worth this group's attention and discussion.

Hill, of the University of Washington and Brookings, began with a discussion of the problems of design and evaluation in the reform of big city school systems. There are, said Hill, many plausible initiatives, but little progress -- for tragic reasons. There is a fair amount known about what is required to have a school work for disadvantaged students, he said. Furthermore, there is some success at reproducing the common characteristics of good schools: Schools have to be very simple and very focused on learning above all other things, and on a centripetal curriculum that is drawing all kids toward a central core of knowledge. Schools that work for disadvantaged students must also be strong organizations that reach out to and socialize parents and faculty as well as students.

What is disappointing, said Hill, is that such schools can be reproduced, but they almost always require special support outside the public system. When such schools are left alone and not supported by foundations or other outside financial and political sources, they usually wither and die. No public school system, he said, has taken responsibility for reproducing such schools in large numbers.

One reason is competition among initiatives. There are large-scale initiatives to retrain all the staff, create community-based parent initiatives, or spend a lot of money on technology. "All these different initiatives become, in a sense, an ecology hostile to one another," said Hill. "And basically what we have is a system in which there is a central policymaking structure in the local school board -- which is willing to let almost

any initiative go on as long as it can survive -- but doesn't regard any one of those initiatives as likely to change the way it does its own business."

As a result, said Hill, unlike a market where there could be a dominant solution that would spread, almost all initiatives last only as long as they have external support and -- whether they work or not -- they die at the end of it. Hill emphasized, as other participants did, that it is impossible to separate evaluation, program design, and change strategies. In public education, it is possible to get permission from school boards and financial support from foundations and others to try almost anything that is plausible, said Hill. All it takes to win that permission is "a good story." There simply has to be a plausible connection between the problem and what you are doing. If kids aren't learning, it must be that teachers do not know how to teach; therefore, retrain the teachers. While that makes perfect sense, so do a lot of other things. Thus, one problem is the gateway -- the requirements for entry as a plausible solution are very low. You do not really have to be able to say how your intervention is actually going to make a change in the long run. On the other hand, the system is built for "serial capture." No one thing can dominate the system, resulting in "a system of continually innovating chaos."

One of the striking findings of his current research on education reform, said Hill, is what he has labeled "zones of wishful thinking." These are areas in which the reform proposed would work only if other important, non-trivial changes in the system occur -- but which the reform itself cannot bring about.

His second finding concerned the dramatic similarities among competing reform proposals. Said Hill; When we shifted our focus, and reasoned not from the initiative to the student but from the student to the initiative, we found that people proposing competing reforms -- everything from teacher training to vouchers -- had strikingly similar ideas about what was a good environment for students: teachers working together with strong, shared ideas about what students should learn, a continuing search for better teaching methods, and parental support. Yet proposals for ways to create these conditions and changes in adult behavior were profoundly different. One set of reformers believe that they were the product of intrinsic factors, such as belief, commitment, and the inherent attractiveness of good ideas. At the other end of the spectrum are those who rely on extrinsic factors, such as rules and regulations, fear of lost employment, and financial rewards and penalties. Few combine the strengths of both approaches.

The third finding Hill reported from his research was in response to the question of why it was so hard for education reformers to reconcile their differences. There are clear economic incentives for reformers to emphasize differences over similarities, but ideological differences over hot button issues seem to be even more important. Many claims are made that actually reflect ideological beliefs more than established facts. ("School choice always creates segregation by class and race." "Competition creates sweatshop conditions for teachers." "Teacher development programs are self-indulgent and wasteful." "Education standards constrain curriculum and impoverish what students learn.") When courts or governors or legislatures look

for guidance on how to reconstruct failing school systems, they generally find only chaos.

In these contentious circumstances, reliable evaluation becomes crucial, Hill concluded. Especially when system-wide initiatives are begun, they must be based on expressly worked-out cause and effect theories, specifying the links between the initiatives and the ultimate goal of improved student learning, so that people can tell in the short run whether what was expected to happen was actually happening.

The Tension between Rigor and Usefulness

Robinson Hollister, Professor of Economics at Swarthmore College, spoke of the current difficulties faced by social scientists and evaluators who must respond to the mismatch between new demands for evaluations of complex interventions that are both rigorous and useful to policy and program people, and traditional methods of evaluation.

The first response has been to lower the standards of evidence -- "a healthy development in same ways." But because credible results depend an the standards of evidence, the big problem for evaluators is “how to back off these high standards… How far do you back off, and under what circumstances?”

A second response has been to increase the use of a theories-of-change approach that has not been widely implemented, but where it has been, its major contribution is that it leads to potentially better program design. Specifically, said Hollister, a theories-of-change approach compels an explicit statement of the relationship between the current activities, ways in which a program will evolve, and ultimate outcomes. On the problematic side of the theories-of-change approach is the problem of the counterfactual. "If we're really concerned about the impact of this program in. the sense, that it changed the outcomes from what they would have been in the absence of the program, these methods do not salve that problem," he noted.

A third response has been to try multiple modes of quasi-experimental evaluations. In other words, mixing in one single evaluation several different modes of quasi-experimental approaches in the hope that there will be confirmatory directions from each of the individual methods. One obvious problem, said Hollister, is that may evaluators think that this would lead to many instances of contradictory, and even wrong, results.

Another very important response – though still not widely used – is the mobilization of small-area data. Hollister said that the evaluation of broad community change initiatives requires small area data that have neither been generally available nor mobilized.

At the other end of the size spectrum there are different evaluation problems. There has been little investment in studying and modeling change at the community level in the absence of any intervention – little knowledge of how communities normally

change. Thus it is difficult to say whether an intervention affected a community. The questions to ask are these: Can we model community change? What are the roles of social networks, social capital, civic entities, and the activities of these communities?

And as Peter Rossi noted, one of the worrisome characteristics of evaluations when the units of intervention are schools and neighborhoods is the difficulty in disentangling within-school or within-neighborhood effects from between neighborhood effects.

One other evaluation method is being used by Public/Private Ventures in some of the communities where it is trying to improve outcomes for young people. That is the use of detailed time-use studies. The idea behind this method was quite basic: If outcomes for young people were going to change, the patterns of their time use would change too.

The Relationship between Evaluation and Theories-of-Change

Peter Rossi, director emeritus of Evaluation Design and Analysis, started with the premise that no evaluation can get at the absolute truth, and described for the group the evolution of evaluations. "All we do is, we make a very, very plausible statement about what the truth might be. That's plausible for a period of time, until the next evaluation comes along, the next experiment conies along which might refute that."

Commenting on the surge of interest in theories-of-change evaluation, Rossi said, "They are not a substitute for effectiveness evaluation." It is important, he said, to understand that implicit models or the program theory on which a program is based are enormously useful to the program, its funders, and other important stakeholders. But knowing the theory is not a substitute for empirical measures of program outcomes and contrasting them with a counterfactual condition.

Attempting to soften his reputation as a pessimist (based in part on his "Iron Law of Evaluation," promulgated in 1987, which holds that "the probability that any given social program will prove to be effective is not high"), Rossi said that most programs can be evaluated for effectiveness and can yield estimates of effect that are convincing. He posited that it is not necessary (although very useful) for a program to hold to a fixed model to be evaluated. In addition, he said, effectiveness evaluation can accommodate nonlinear program models. There is no reason why we cannot have a nonlinear response form.

Rossi observed that there are many ways in which evaluation design and analysis are changing and evolving. First, the research community is learning better how to handle massive data collection and massive data analysis, which is helpful in overcoming the mismatch in time between how long it takes to do an evaluation and how long a policy issue is on the agenda. As an example, he cited the early massive income-maintenance experiments of the 1960s and 1970s, which took years to produce results. By the time the results appeared, the issue of income maintenance was no longer

part of the political debate. Today, he said, there is a much closer fit between policy time and evaluation tinge, citing the work of MDRC on the waivers allowed under the old AFDC program, which produced results in what he called a "remarkably short period of time," defining that as less than a decade or less than five years.

Rossi also lauded the trend of some evaluations to blend intelligent and sensitive combinations of different approaches, citing as an example the current effort of the Urban Institute's New Federalism Project, which will combine traditional evaluations with rich descriptions of how public welfare agencies of 13 states react to welfare reform. He also noted that in some areas, there have been enough evaluations of particular topics that we can learn from systematically going over those evaluations and extracting from them their commonality. "It's not perfect knowledge," he said, "but it's good knowledge." And finally, Rossi noted there are a few examples -- mostly in the public health field of randomization -- using social groups larger than households as intervention units.

Responding to Tom Mann's question about the relationship between theories-of-change and evaluation, Bill Dickens of the Brookings Institution argued that while randomized experiments are highly desirable for evaluating policy interventions, there is a more substantial basis than is usually acknowledged for a rigorous use of theories-of-change. He noted that recent work in the history of science suggests that the classic scientific method does not capture the rich tapestry of actual scientific practice, which admits a wide range of evidence. Broadening one's methodological perspective does not necessarily entail lowering one's standards nor abandoning rigor.

Dickens outlined a strategy of using Bayesian statistics, with its central role for prior probabilities to explicate and evaluate an expected sequence of outcomes. While this approach might face formidable obstacles in the field, it offers the promise of using theories-of-change to learn about the efficacy of complex policy interventions without sacrificing scientific rigor.

Mann then asked Hollister to comment on the implications and risks – in a less than perfect world – of proceeding with evaluation techniques adjusted to deal with complex interventions. Will they produce useful and reasonably reliable information? Or are we deluding ourselves and finding more success than actually exists?

Hollister responded by saying that it is still too early to tell because more of the newer evaluation efforts are still in process. But, he cautioned, even when you develop a solid and sharp theory-of-change and have thereby strengthened program design, you are still left with the problem of the counterfactual.

Iterating toward Change

Orr pointed to another shortcoming of the theories-of-change approach to evaluation: It will only support causal inference to the extent that it can make fairly precise predictions that can then be verified or disproved by observation. Most social

theories allow at most qualitative predictions. Orr contrasted this state of affairs to the physical sciences, which originated the "scientific method" of advancing knowledge by stating a theory and then attempting to confirm or disprove it. For example, in the early 1940, physicists predicted that if the big bang theory of the universe were correct, we would observe a "cosmic microwave background radiation" with a temperature of about three degrees Kelvin. "Forty years later, we put up a satellite and measured the temperature of that radiation at 2.73 degrees Kelvin," he said, "Now that's a precise prediction!" Most of the predictions for social interventions are much less precise: "If we change this feature of the public schools, performance will go up."

Moreover, in social interventions, the process of change – and the process of defining a theories-of-change – is an iterative one. "And unfortunately, as you iterate, the programs go along," said Orr, "so the pitfall here is that the theory keeps evolving until it is really nothing more than an ex-post-facto rationalization for whatever happened." He also noted a second danger that evaluators become so involved that they become cheerleaders for the programs they are evaluating.

Promising Practices and Next Steps

The participants then turned to a consideration of the new work necessary to address the issues that surfaced in the earlier discussion.

Robert Granger of MDRC suggested that a theories-of-change approach should advance most of the purposes of evaluation if evaluators took seriously the issue of the counterfactual and did the following four things: developed strong theories-of-change; used multiple methods of inquiry to search for already confirmed patterns and results; creatively blended designs; and were patient about replication.

In considering how to develop strong theories-of-change, he said, "The first question is theories about what?" And he suggested a need for strong theories about implementation, innovation, and change, before one even reaches the question of effects. He noted that most evaluators are naïve about synthetic literature on how organizations or communities change or how they respond to interventions. Granger said that while there are good data about how individuals change over time, there are not good data on how families or organizations— or small areas or places – change. He also pointed to the need for good theory about the near-term, intermediate outcomes that we ought to be paying attention to because they might dramatically relate in a predictable fashion to valued long term outcomes. Examples would include the concept of collective efficacy coming out of Rob Sampson's and Tony Earl's work in Chicago, and the student engagement with schools that Jim Connell's work suggests is likely to be an important predictor of improved school achievement.

Strong theories would also pay attention to how race, gender, class, and other contextual factors interact with interventions. Most models, Granger noted, include such considerations, but they are "at best out to the side, sort of arrows coming in from the

heavens." If these issues were on the table, we would be a lot closer to being able to use theories-of-change effectively.

As for using multiple methods to search for patterns, Granger said that most studies are not very strong on integrating quantitative and qualitative methods. At best, they attempt to confirm findings through a variety of means and simultaneous events. And that, he said, is not a good integration. He urged participants to think about how to integrate methods in a more "planful" way.

Granger also said it was important to use multiple methods to measure the "penetration" of various interventions into the daily lives of individuals, organizations, and communities. "We have a lot of presumptions about 'if we do this,' then it is going to be felt in some demonstrable way, within families, or within schools, or within whatever." Those presumptions need to be tested, and new thinking is needed to study the relationship between participation and dose, and what "participation" in a community initiative means.

In remarks about blending designs, Granger said that when evaluators do not use the units in which they are interested as their own controls (such as time series), they run into a problem of selection bias. They must worry about whether one community is like another in all the relevant characteristics if they plan to use one for an estimate of the counterfactual. Thus a lot of designs try to use a cross-cohort or times series approach instead, using previous history as the counterfactual. And then, Granger said, they "run squarely into the problem of contemporaneous historical events discombobulating things."

Turning to replication, the basic problem, said Granger, is whether it is possible to create the patience and taste to wait, and to avoid being captured by the rush to judgement and urgency that people feel in the field. One of the reasons that the work-welfare studies in the 1980s had some real power, he noted, was because there were similar experiments in eight different locales with similar outcomes. "If we are going to seriously understand something, we're going to understand it because we have looked at similar interventions across a variety of places and a variety of times."

Today the question for evaluators is how to describe various interventions and measure those processes in common kinds of ways. If someone is trying to empower a groups of people within a community and they are using particular strategies to do that, said Granger, "you'd like to be able to measure those strategies in at least reasonably consistent ways," and to measure empowerment, whether it happened, and whether you got more or less of it.

Gary Walker of Public/Private Ventures began his remarks by pointing out the need for a new intermediary that could more systematically collect and analyze information about comprehensive community initiatives. He asked the group to consider that the need for broader experimentation, for more support of comprehensive initiatives, may be more urgent than the need for addressing the technical problems of evaluation.

"What are these things we're talking about?" he asked. "What clarity can we bring to what a comprehensive community initiative is?" Walker said that the typical comprehensive community initiative that fails to bring about comprehensive community change is "one and a half staff people in a community of 40,000 with a poverty rate of 42 percent and an adult unemployment rate of 28 percent." The probability of anything coming out of most of these efforts is so small that it would be a shame to waste resources evaluating them – and a shame to end up with another stack of evaluations that shows that nothing works.

He suggested that the only honest way to evaluate comprehensive community initiatives is that they be tested "adequately and on a large scale and not think we can do them in a couple of small neighborhoods and then extrapolate from there."

The second key issue beyond methodology for Walker is the issue of asking the right questions – from a political point of view. Will evaluations provide information on the practical things that skeptics will find satisfying and useful? Walker noted that 10 years of research on national service – costing perhaps as much as $13 million – have not answered the question most likely to be asked by conservative critics: "Does national service promote volunteerism?"

In the case of comprehensive community initiatives, Walker suggested two outcomes that must be defined and measured carefully: empowerment and local involvement. And from a political perspective, it may be important to limit the use of the term comprehensive community initiative and to limit expectations. He pointed to P/PV's work in that neighborhood; increase the amount of after-school and weekend activities; and increase young peoples' belief – early on – that success in work is connected to success in school.

Sawhill, echoed by Barbara Dyer of The Public's Work, raised the issue of understanding changes that occur naturally. Sawhill said these are particularly significant when evaluators compare time series sets of changes to the usual kind of cross-sectional evidence from program evaluations. She offered the example of teen sexual activity, noting that sexual activity rates among 15 year olds have tripled in 25 years. Yet numerous deliberate attempts to modify sexual behavior have either been unsuccessful or have had minute effects.

"Why is it," Sawhill asked, "that there's been such a huge change and we can't somehow or other take advantage of what's caused that change and design a program around whatever we think that thing is?" Sawhill hypothesized that the discrepancy suggests that either the interventions are much too timid or too short-lived. Dyer suggested and Sawhill strongly concurred that the reason is probably a major shift in social norms. "I really think we've neglected – in both program design and program evaluation – this whole area of the change in social norms, which goes usually with changes in behavior."

The Fuzzy Line between Design and Evaluation

Janice Molnar of Atlantic Philanthropic Service Co. pointed out that Sawhill's questions highlighted the fuzziness of the boundary between program design and program evaluation. "What is evaluation anyway?" she asked, "Is it measure impacts or is it working with program managers and architects around the design of an appropriate intervention?" Several participants agreed there was a tension between program design and evaluation, pointing out that evaluators walk a fine line when they try to provide useful feedback for program changes because programs are often afraid of the evaluator as someone who stands between them and the funder.

Others indicated they were not troubled by the lack of clear distinction between program evaluation and program design. Each should continually inform the other. Granger suggested there was no conflict between helping programs do a better job of creating a program, and rendering some bottom-line judgement on the effectiveness of that program when it's done. When considering comprehensive community initiatives, he said, "you have to think in the same way regardless of whether your purpose is purpose A or purpose B." As an example, he said that MDRC, which is currently working on the Jobs Plus demonstration, is spending a lot of time trying to figure out how to help sites create something that is consistent with a set of principles, but that may look different in each particular manifestation.

Schorr indicated that it was important to shape evaluation in ways that allow it to support good program design. She agreed with Sawhill's contention that the forces that seem to bring about major changes are far less circumscribed than our interventions -- in part because evaluators push program people to circumscribe their interventions, and both evaluators and administrators push program people to minimize variation. "Whenever you go any place and observe that effective programs tend to vary from community to community, both the administrators and evaluators turn very pale," she observed.

Ron Ferguson of Harvard University's Kennedy School of Government raised the issue of lack of uniformity from one site to another in another context, citing the Comprehensive Employment `training program. He noted that the evaluation results from the West Coast indicated the program made big differences for clients. But in the replication process, East Coast sites adopted the name, but very few of the practices. And while some involved in the program suggested that no one had the right to insist that each site replicate specific practices in the program, Ferguson said that evaluators have an obligation to take note of these differences.

When he evaluated YouthBuild, Ferguson said he also identified the necessary components of YouthBuild and made it clear in the evaluation that some programs were using the YouthBuild name, but did not have all the components. Speaking of these programs, Ferguson raised a second factor for evaluators to take into account. One of the big differences in effectiveness has to do with the talent of the people in lead positions. Part of what we evaluate is our notion of what makes an effective staff person. What kind of training and background do they have? What kind of philosophical orientations are

required? He suggested, “If we could get to the point where we say we aren’t sure exactly what they’re going to do, but we know if we hire somebody to fit this profile and stick them over there and tell them to pull people together to problem-solve, then five years later when we turn around, they will have solved some problems.”

Reflecting back over the day, Schorr suggested that while the discussants had reached few unanimous conclusions, the participants had generally agreed that it is possible to learn more systematically from the experiences of complex social interventions. “And maybe if we did this conference over again, we might call it systematic learning about complex social interventions,” she concluded.

Appendix A: Participants

Henry Aaron
Brookings Institution

Geraldine Brookins
W.K. Kellogg Foundation

Gary Burtless
Brookings Institution

Richard Darman
The Carlyle Group

William Dickens
Brookings Institution

Barbara Dyer
The Public's Work

Ronald Ferguson
John F. Kennedy School of Government

Keith Fontenot
Office of Management and Budget

William A. Galston
University of Maryland

Robert Granger
Manpower Demonstration Research Corporation

Margaret Hamburg
Office of the Assistant Secretary for Planning and Evaluation
Department of Health and Human Services

Paul Hill
University of Washington

Robinson Hollister Jr.
Swarthmore College

Anne Kubisch
Aspen Institute Roundtable on Comprehensive Community Initiatives

Thomas E. Mann
Brookings Institution

Janice Molnar
Atlantic Philanthropic Service Co.

William A. Morrill
Mathtech, Inc.

Larry Orr
Abt Associates

Peter H. Rossi
Evaluation Design and Analysis

Isabel V. Sawhill
Urban Institute

Mark Schmitt
Open Society Institute-Washington

Lisbeth B. Schorr
Harvard Project on Effective Interventions

Stephanie Shipman
General Accounting Office

Kathleen Sylvester
Social Policy Action Network

Gary Walker Public/Private Ventures

R. Kent Weaver
Brookings Institution