

—EDITED TRANSCRIPT—



HUDSON INSTITUTE'S

## BRADLEY CENTER

FOR PHILANTHROPY AND CIVIC RENEWAL

*presents*

# Metrics Mania?

Thursday, March 20, 2008 ▪ 12:00 to 2:00 p.m.

Hudson Institute ▪ Betsy and Walter Stern Conference Center ▪ 1015 15th Street, NW ▪ Suite 600

In the early 1970s **GARY WALKER** established a jobs program for ex-convicts and recovering drug addicts. Within a few years it was expanded to nine other locations and evaluated as a national demonstration. The evaluation showed that the program did not substantially change the lives of those who participated. This was only the beginning. As the “Evaluation Revolution” unfolded over the next thirty years, the overwhelming majority of social programs with impact studies would be shown to have similar results: no substantial change in the lives of participants. In an essay prepared for this discussion (“Reflections on the ‘Evaluation Revolution,’” attached) WALKER, former president of Public/Private Ventures, describes “a sophisticated and maturing field of evaluation thrust upon an ever-immature field of demonstration social programs.”



Why is it that philanthropy has learned so much about metrics and yet has so little by way of measurable success to show for it? On March 20, Hudson Institute’s Bradley Center hosted WALKER and other experts who have grappled with measurement, including **KATHLEEN ENRIGHT** of Grantmakers for Effective Organizations, The Meyer Foundation’s **ALBERT RUESGA**, and **HOWARD ROLSTON** of Abt Associates and The Brookings Institution. The Bradley Center’s own **WILLIAM SCHAMBRA** served as the discussion’s moderator.

### PROGRAM and PANEL

- 12:00 p.m. Welcome by Hudson Institute’s **WILLIAM SCHAMBRA**  
12:10 Panel discussion  
Panelists: **GARY WALKER**, formerly of Public/Private Ventures  
**KATHLEEN ENRIGHT**, Grantmakers for Effective Organizations  
**ALBERT RUESGA**, The Meyer Foundation  
**HOWARD ROLSTON**, Abt Associates and The Brookings Institution
- 1:00 Question-and-answer session  
2:00 Adjournment

THIS TRANSCRIPT WAS PREPARED FROM AN AUDIO RECORDING and edited by Krista Shaffer. To request further information on this event or the Bradley Center, please contact Hudson Institute at (202) 974-2424 or send an e-mail to Krista Shaffer at [Krista@hudson.org](mailto:Krista@hudson.org).

## HUDSON INSTITUTE

1015 15th Street, N.W.  
Suite 600  
Washington, DC 20005

202.974.2400  
202.974.2410 Fax  
<http://pcr.hudson.org>

## PANEL BIOGRAPHIES

**Kathleen P. Enright** is the founding executive director of Grantmakers for Effective Organizations (GEO). Enright speaks and writes regularly on nonprofit and grantmaker effectiveness. She serves on the board of the Fieldstone Alliance and advisory board of The Center for Effective Philanthropy. Previously, Enright has worked as the group director, Marketing and Communications, for BoardSource and served on Independent Sector's Building Value Together Committee.

**Howard Rolston**, a highly regarded social researcher who changed the way federal programs for children and families are evaluated, joined Abt Associates in August 2006 as a part-time principal associate. In addition, since 2004 Rolston has held positions at The Brookings Institution, where he continues to consult as a visiting scholar. Prior to joining Abt and Brookings, Rolston was director of Planning, Research, and Evaluation at the U.S. Department of Health and Human Services' Administration for Children and Families, with responsibility for designing and funding numerous major experimental evaluations. In addition to his government service, where he won four Distinguished Service and Meritorious Executive awards, Rolston also served as an assistant professor of philosophy at Georgetown University.

**Albert Ruesga** is vice president, Programs and Communications, at the Eugene and Agnes E. Meyer Foundation. He joined the Meyer Foundation in October 2001 from the Forum of Regional Associations of Grantmakers, where he served as the founding director of the national New Ventures in Philanthropy initiative. Prior to that, Ruesga was the donor resources manager at the Boston Foundation, where he provided individualized philanthropic services to major donors and donor prospects; staffed special funds and initiatives; and designed introductory grantmaking courses for advisors and young professionals. He serves as chair of Hispanics in Philanthropy and as treasurer of The Communications Network. He also serves on the boards of the Washington Regional Association of Grantmakers and the Catalogue for Philanthropy—Greater Washington; as an advisor to Caring for Change; and as a volunteer instructor for The Grantmaking School based at Grand Valley State University. Ruesga is an accomplished author of articles in *Social Theory and Practice*, *The Journal of Popular Culture*, and other publications.

**Gary Walker** is president emeritus of Public/Private Ventures; he served as P/PV's president from 1995 through May 2006. A former Wall Street attorney, Walker began his social policy work in the 1970s at the Vera Institute of Justice, where he set up Pioneer Diversified Services, Inc., a New York City-based business that hired only ex-offenders and recovering substance abusers. He then managed the National Supported Work Demonstration, one of the earliest attempts to make work a key element of our nation's welfare and social service policies, in his capacity as senior vice president of the Manpower Demonstration Research Corporation (MDRC). Walker went on to direct the first national study of the Job Training Partnership Act, the nation's major policy vehicle for job training during the 1980s. He joined P/PV in 1986 to pursue that goal, providing leadership for initiatives in mentoring, after-school programming, violence prevention and workforce development, and for work with faith-based organizations, national service and volunteer programs.

# HUDSON INSTITUTE WHITE PAPER

## Reflections on the “Evaluation Revolution”

by Gary Walker

I LEFT my job as a Wall Street lawyer in the early 1970s to establish a jobs program for ex-convicts and recovering drug addicts in mid-town Manhattan. Like many of my generation who created or ran social programs, I went into this endeavor with the optimistic assumption that a social program would be the door to significant change in the participants’ lives: no more crime, no more addiction.

Coming from Wall Street, I was familiar with bottom lines, hard outcomes and mounds of data to prove or disprove efficiency and viability. But the cause of helping others made those terms fade in importance, and though the Department of Labor provided funds to my project for collecting demographic and performance data, the human and social cause aspects of the job—plus the more mundane aspects of operating an organization—took precedence. A formal evaluation would have been an annoyance to me—at best.

I don’t think my attitude towards evaluation was or is unusual for a program operator. And in the 1970s most foundations were sympathetic to it; they were either about helping individuals or fostering “social change.” They were not eager to put their resources into costly evaluations, which would do neither.

In the three-plus decades since those early, enthusiastic and optimistic days of social programming, the most consistent and notable trend has been the dramatic increase in and enthusiasm for formal evaluations, in both the public and philanthropic sectors. No matter whether the administration has been Democrat or Republican, liberal or conservative; the foundation old or new, the donor deceased or in charge: evaluation has become a critical element of all social initiatives.

---

This Hudson Institute White Paper was prepared for “Metrics Mania?” a panel discussion hosted by Hudson Institute’s Bradley Center for Philanthropy and Civic Renewal on March 20, 2008.

## Reflections on the “Evaluation Revolution”

by Gary Walker

“Effectiveness must become the principal criterion for givers of time and money.” This clarion declaration is the first of five conclusions of the 1997 report of The National Commission on Philanthropy and Civic Renewal, *Giving Better, Giving Smarter*.

The national office of United Way—whose local chapters raise over \$3 billion—initiated several years ago a major project to both emphasize the importance of specifying outcomes for local giving, and to provide assistance to local chapters on how to go about determining the outcomes of individual grants. This effort, funded by the W. K. Kellogg Foundation and the Ewing Marion Kauffman Foundation, resulted in *Measuring Program Outcomes: A Practical Approach and Focusing on Outcomes*.

Grantees report that never before have grant negotiations with foundation staffs been so focused on specifying outcomes. Some foundations have employed consultants to work with their staffs so that input, operational processes, and intended intermediate and long-term outcomes and impacts are specified and differentiated. Many have added evaluation departments to their organizational structure. Small and medium sized foundations, which have previously given exclusively to direct services, are now asking for and funding evaluations, so that they may know with objectivity and rigor if the projected outcomes are achieved.

The list of declarations and initiatives to increase the philanthropic resources and efforts that go into evaluation is near endless. Just this morning (February 15) I talked with a recently hired staffer at the Gates Foundation, whose job is to set up a new department, called “Impact Planning and Improvement,” which will work with all the substantive departments of the Foundation in developing theories of change and results measurement strategies.

So you might say, what’s the problem? The field of social programming is growing up, maturing, and developing a sense of responsibility beyond the rhetoric of social change and the rosy glow that comes from being in the business of helping others. For some, that ends the discussion: the search for measuring effectiveness joins the ranks of motherhood and apple pie, not to be questioned.

And that perspective achieved a direct hit on me: the jobs program I ran was in the mid-1970s expanded to nine other locations, and evaluated as a national demonstration. The results were displayed on elegant tables, with asterisks and footnotes and numbers to the third decimal, but in the end the results could be expressed in simple narrative: the program did not substantially change the lives of the ex-convicts and recovering addicts who participated. The control group, in terms of jobs, income and avoidance of prison or substance abuse, succeeded at about the same rate as the participants.

What I had learned by then is that it is not so easy to change lives. The program I ran was okay, but the quality of its services, and the depth of change it provided, were nothing extraordinary. As the results showed.

I spent the next three decades working for organizations that conducted large-scale national research demonstrations (MDRC and Public/Private Ventures); the evaluations were the critical

## Reflections on the “Evaluation Revolution”

by Gary Walker

element of these demonstrations, the *raison-d’être* of the organizations. It is that experience which produced other lessons, lessons that complicate the evaluation storyline. These lessons need not dampen our enthusiasm for effectiveness and accountability; but they do illuminate factors that affect whether and how we evaluate, and if there are other more pressing needs.

Lesson 1: Outcomes are not impacts. New philanthropists often confuse the two, because almost every social initiative declares outcome goals: in jobs programs, place 75 percent of participant in career-path jobs; in education, graduate 90 percent of freshman participants; *etc.*

It turns out, though, that not only is there little congruence between outcomes and impacts, but that the general lesson from the past three decades is that the higher the outcomes a program has, the less likely it is having an impact.

A good example is the 1983 federal Job Training Partnership Act (JTPA). Its programs aimed to place poor people with multiple obstacles to employment into jobs; the Act put a major emphasis on quantitative placement rates as a measure of local success in achieving the act’s goals. Local administrators set very high goals for the programs they funded, and offered financial incentives. As local placement rates around the country began to soar—to over 80 percent in many locations—and were verified as factually accurate, critics speculated that these rate were *too* good, and indicated that most JTPA participants did not have serious obstacles to employment and would have gotten jobs even without JTPA’s modest training interventions.

JTPA advocates scoffed. Then an impact study was done. It basically supported the critics. Well-specified outcomes, careful measurement, incentives and good performance all amounted to very little added value.

Why does this occur? Primarily for two reasons. First, it is not as easy as it appears to predict who is going to fail in school and in life. Sometimes advocates would have us believe that all poor people, all illiterate people, all children of one-parent families, are going to end up jobless, in prison, on drugs or in some other desperate condition

But that is not true. The *odds* are higher that they will—but many, sometimes the majority, will, on their own, do all right in life.

So programs which use broad categories of disadvantage—as most do—are taking in many people who would have done all right without the program—and when the evaluation includes a control group, the results will show that.

Second, the higher the outcome goals, the more a program will try to select from the eligibles those who seem more likely to succeed—how else can high goals be reached? But a well-constructed impact evaluation will include in the control group only those that the program would have accepted—so they too will succeed at a high rate.

Thus evaluations that only track outcomes, and do not include a control group, often lead to satisfied funders and operators—but in fact are achieving no more than would have happened anyway.

## Reflections on the “Evaluation Revolution”

*by Gary Walker*

This now well documented phenomenon has led many to conclude that an impact study, with a carefully constructed control group, is the only kind of evaluation worth doing. Thus the last 20 years have seen a significant increase in impact studies, and the growth of an industry with a number of strong organizations who can carry out large-scale multi-site impact studies. The technical expertise involved is impressive, and the mathematical techniques involved are well beyond the intelligent layman.

In short, the science of evaluation of social programs has made enormous progress in telling us whether social programs make a lasting difference in participants’ lives. The result?

Lesson 2: The overwhelming majority of social programs with impact studies do not show a significant change in participants’ lives a year or two after the program. This phenomenon has not changed over the thirty-odd years that social programs have been evaluated; that is, while the science of evaluation has been improving and growing, in sophistication, size and resources devoted to it, our ability to actually improve lives through social programs has been consistently unimpressive.

This disjuncture in progress is not entirely surprising; the human ability to make technical progress has throughout history always surpassed our ability to make progress on less orderly social issues. But the disjuncture does belie the earlier conclusion that the increased interest in evaluation is a function of a maturing field of social programming. What is maturing is the field of evaluation; the field of social programming seems to be stuck in the same no-to-low performance mode that it’s been in for the past forty years.

If you stopped at these two lessons, you might conclude that the “evaluation revolution,” or “outcomes movement” as it is more popularly misnomered, has performed a great if discouraging service: it has shown us that outcomes per se are a deceptive measure; that setting high outcome goals does not necessarily increase impacts; and that social programs don’t accomplish much. You might take a bold step further and conclude that social programs are a waste of money, whether the taxpayers’ or philanthropists’.

The conclusions regarding outcomes are well founded, and are a great service. The conclusions regarding impacts are complicated by another lesson:

Lesson 3: The weak impacts we’ve evaluated are in good part an artifact of our approach to setting up impact studies. Impact studies are expensive; require significant numbers of participants to be highly reliable; and often are controversial, since control group members cannot receive the programs’ services.

This combination of requirements means that government and large foundations are the likely funders; they have the money. Both these sectors hold innovation in high regard, in contrast to building on “old” ideas, or the ideas of competitors, and thus want to test and evaluate innovations. Multiple locations to test the ideas must be found, in order to achieve the high numbers required. New operations are often required, both to get the numbers, and to avoid the controversy over control groups, which is usually highest in established program operations.

## Reflections on the “Evaluation Revolution”

by Gary Walker

The result of all the above is that the majority of well-done impact studies are on relatively new operations, carrying out relatively new ideas. And if the study contains an implementation evaluation component, there are almost without exception two major findings: that the new idea was not well or consistently implemented across the multiple sites; and that new operations had significant “startup” problems.

So it is usually not clear whether we’re measuring a bad idea, or simply the poor implementation of an idea. This is a shaky foundation to come to the conclusion that social programs are a waste of money—especially when a small group of impact evaluations on established programs, with decades of experience and high implementation standards and fidelity, such as Big Brothers Big Sisters and the Nurse-Family Partnership, have produced some impressive impacts.

Plus we simply don’t know about the impact of a whole class of social programs—small, community-based operations that have too few numbers to ever be part of an impact study.

The “artifact” lesson is hardly the fault of evaluators; they evaluate what funders want evaluated. But it should make us pause in our rush to be accountable by means of impact evaluations. For the implication of this lesson is that the first things funders need to be accountable for is the quality of the program which they’re funding. That requires patience, and a use of funds for things like training, and the development and implementation of quality standards, indicators and tools.

This lesson is not new. In the late 1970s I traveled to England to look at some of their social programs; they were stunned that we did impact evaluations without being *sure* that the program was well implemented. And in the mid-1990s our own Department of Labor did a review of all the evaluations it had funded, and concluded:

It often takes time for programs to begin to work. Many of the success stories for the disadvantaged have come from programs which were operating for five years or more before they were evaluated. (U. S. Department of Labor, 1995)

The problem is that our impatience for results, the fact that new political administrations and new philanthropists want to be known for their innovations—means that it is not a lesson well learned. Instead these factors conspire to promote a sophisticated and maturing field of evaluation, thrust upon an ever-immature field of demonstration social programs.

Some evaluators are aware of this dilemma, and warn funders and innovators, and insist upon assessing quality before proceeding to impact evaluations. But it seems unrealistic to expect the evaluation industry to correct this practice: it needs work to satisfy its own successful growth, and the most lucrative work comes in the form of large-scale, multi-site research demonstrations. And there are always new political administrations, and wealthy new philanthropists, to satisfy that need.

The “evaluation revolution” has produced a number of important insights, most notably the counterintuitive, often inverse relationship between outcomes and impacts; and the fact that most

## Reflections on the “Evaluation Revolution”

*by Gary Walker*

programs are not well implemented, and thus don't produce much in the way of impacts. Those that do produce impacts tend to be well established, with clear standards of quality.

The challenge for philanthropy and the public sector over the coming decades relates less to measuring outcomes or impacts, or to churning out innovations, than to helping establish—and measure—quality in program performance. It will, I think, be a hard challenge to imbed in funders' practice, as it does not have the excitement of innovation, or the elegance and certainty of impact evaluation. And in an impatient culture, delays our need for immediate results and accountability.

But its advantages would be the actual maturing of the world of social programming, and those that fund it. The emphasis would shift from creating answers at the political or philanthropic level, and then further creating demonstrations to test those answers, to looking at the thousands of efforts already created locally, and helping the most promising with the resources and assistance necessary to achieve quality. If some seemed unique enough to replicate, the resources and time to do so with quality would be the standard—not doing it so fast that an impact demonstration can be set up right away.

But what about results, actual impacts? We just wouldn't know within the usual five years. Maybe not for ten or fifteen years. But then we've spent the better part of four decades learning that actual impacts are very hard to achieve, and definitely can't be achieved without quality implementation.

Progress often involves slowing down, reflecting on the path taken, and shifting the course. I think that's the spot we're at in the world of social programs, and in our thinking that a focus on evaluating results will in itself produce better results.

## PROCEEDINGS

WILLIAM SCHAMBRA: Good afternoon! I'm Bill Schambra, director of the Bradley Center for Philanthropy and Civic Renewal here at Hudson Institute. Krista Shaffer and I welcome you to today's panel, entitled "Metrics Mania?"

First, a note about our next panel. For April, we've chosen what is proving to be a particularly provocative topic. We will examine the question of diversity in philanthropy, anchoring the conversation in a look at AB 624, the legislation recently passed by the California State Assembly requiring the major foundations in the state to report on their own composition and that of their grantees and contractors according to race, gender, sexual orientation, and so forth. Joining our panel on April 7 will be John Gamboa, executive director of California's Greenlining Institute, which did the research behind AB 624; Pablo Eisenberg, well known to all of you as an insightful critic of foundations; Heather Richardson Higgins, on the board of the Philanthropy Roundtable; and Renee Branch, director of diversity and inclusive practices at the Council on Foundations.

Now for today's panel. For some of you, March 20 can mean only one thing – March Madness. This is, after all, the day and almost the moment that the NCAA basketball championship tournament kicks off. But for us here at Hudson, March 20 means not "March Madness" but "Metrics Mania" – something different, but perhaps not all that different. At this point, a sign should appear beneath the podium that says, "Warning: Sports analogy to be attempted by non-authorized personnel." After all, as we think about philanthropy's mounting desire to ensure that its outcomes are certain and measurable, aren't we facing the same human impulse that we see reflected in the famous sixty-five-team brackets so much in evidence on office computers today, the yearning to be able to say with "measurable certainty" that this team is better than that team, and that this team finally is the best of all? Philanthropy is perennially frustrated because it can never quite get there. Its programs continue to operate instead in the mode of NCAA football – oh, yes, we're not done mangling this analogy quite yet (laughter) – where in spite of the bowl championship series, there is still an immense amount of subjectivity and guesswork – or perhaps it's better to call it informed judgment – in the decision about who is better than whom, and finally, who is number one.

There are, of course, those who would suggest that the very nature of philanthropy forbids it to get beyond the NCAA football to the mathematical precision of NCAA basketball. At any rate, to help us sort all of this out, we asked Gary Walker, who has just retired from a long and distinguished career at Public/Private Ventures (P/PV), to write for us a white paper bringing to this discussion his best thoughts and his great experience on this topic. Joining him today on the panel will be Kathleen Enright, founding executive director of Grantmakers for Effective Organizations (GEO), now celebrating its tenth year. Congratulations, Kathleen.

KATHLEEN ENRIGHT: Thank you!

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

WILLIAM SCHAMBRA: We also have Howard Rolston, currently at Abt Associates but known to anyone in the social policy field for his many exemplary years at the US Department of Health and Human Services, where he greeted generations of youthful program planners who assumed that they had just invented a new and revolutionary policy with the deflating observation that we had already tried exactly that approach in 1976, 1982, and 1994, and it had failed every time. And finally, we're pleased to have with us for a return visit Albert Ruesga, vice president for programs and communications at the Eugene and Agnes Meyer Foundation here in town. So, Gary –

GARY WALKER: Sure! You know, Bill, it's also the first day of spring; that's another way to look at it.

(Laughter, cross talk.)

GARY WALKER: I just have to take two of my fifteen minutes today to tell you that when Bill (Schambra) called and asked me to do this – you know, when you get into your mid-sixties, as I have, you get these calls asking you to reflect on things. With the first one, you're just honored because it's a reflection of wisdom, you know. But with the second and third ones, it occurs to you that it's really because you're *old*. (Laughter.) I mean, how many people have been around that long?

So I enjoy doing this and all, but the oldness aspect is always a part of it, and I had been feeling a little down about that as you occasionally can feel. And as I was packing for this trip, my wife said, "You always carry that heavy bag. You know, you're getting older. You should use one of these things with wheels on it." And since I was in a slightly weakened state, my pride and confidence, when she offered to lend me hers I said, "Okay."

As I got out of the car and tried to walk to the Philadelphia train station with this bag rolling behind me, something I'd never experienced before – and it was pretty terrific, I walked off the curb and one of the bag's wheels hit before the other, and it twisted. And I tried to keep ahold of it, and it nearly broke my wrist. (Walker raises his bandaged wrist.) And I fell on the ground. And as I fell on the ground, I must have said something like, "Damn this newfangled thing!" And some twenty-year-old who was walking by helped me up, and in perfect innocence exclaimed, "You must really be old if you think wheels are new." (Laughter.)

So I come here with great humility about reflections on anything today. (Laughter.) But I'll jump into it anyway.

When I left Wall Street in 1969 and went into this social policy programming work, running a program in the middle of midtown Manhattan for ex-convicts and ex-addicts, the foundations and the federal government who were putting money into my program sent someone out to sit in my office and talk to me every six months about how things were going. Even I recognized that it was kind of a foolish way to evaluate how all of this money was being spent. They did not really walk around and talk to anybody else. There was very little data on how it was going. So when the program that I was running was – Ford and the US Labor Department and Department

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

of Health and Human Services were going to turn it into a national demonstration, and they were creating at that time MDRC (founded as the Manpower Demonstration Research Corporation) in order to evaluate it, I went over there because I thought, gee, this is a good thing! Coming from Wall Street where everything has a number to it, I had thought that this was certainly a part of the world where nothing has a number to it. It's all heart and commitment and all of these important things. But it's a good thing to find out whether it really works.

And so I have spent a considerable part of my career with MDRC and P/PV evaluating things. But as I look back over this enormous number of years – over thirty – I would say that the trend towards evaluation and trying to know whether something works or not has been the single biggest and most consistent trend in the philanthropy and social programming and policy. Certainly no substantive trend has been consistent – and that's a function in part of democracy and its politics, and the fact that all of these new foundations came on the scene. New things happen all of the time, go up and down. The one trend has been the building and development of an evaluation industry, and particularly one that can perform randomized, control-group impact evaluations on what are perceived to be important social programs or policies. That's the single biggest trend. It has almost become like motherhood and apple pie. How can you really be against wanting to know whether what you did worked or not? And if you look, now, in federal regulations, in federal law, in administrative regulations at the state level, and at foundations and the people they hire, evaluation is now a big part of everything. And as I said, you can't really on the face of it come out and be against it – although I'll be interested to hear if someone tries it. Particularly in a field like this, where there is so much emphasis on commitment and your heart and all, and that's great. But at some point it's important to know whether you're accomplishing anything.

Now I have, myself, wondered, as modern science keeps pointing out that redundancy is in the nature of all biological and other systems and is a secret to their success, why is it that in the field of helping others we have gotten so strict about making sure that we're absolutely getting great outcomes or you're out the door. It does seem to me that we don't do that anywhere else. But that's an aside. Still, you have to ask the hard question.

And I do think, as I look back over thirty years – or more, there have been some really great contributions from this. One is simply the contribution of understanding that good outcomes are not the same as impact. For many in this room – most – it seems like kind of an old lesson that came out of some of the very first demonstrations in the 1970s. As we have new leaders coming into both philanthropy and the public sector, it's a lesson that I think continually needs repeating. We certainly, I think, saw it exemplified at the highest levels in the 1990s with the evaluation of the Job Training Partnership Act. For those of you who don't know, the work that went into establishing an outcomes orientation to the Job Training Partnership Act was enormous – in setting up incentives – all designed to make sure that those job training programs which were going to take in underprivileged people got high outcomes. And you know what? They kept getting better and better outcomes. In fact, their placement rates in most jurisdictions got over 80 percent. And then there was an impact study, and what did it say? The control group was getting jobs at a rate of over 80 percent, too. As in most human things, by simply setting up outcomes as your goal, the institutions in charge figured out a way to get there, and if there was an easier way

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

than doing really hard work, like namely finding those people in the population who'd be easiest to deal with, they did.

So that, to me, out of what I call the "evaluation revolution," is a really, really important point. Even the new group of philanthropists coming up, as sophisticated as they are and the businesses that they've run, are actually just used to outcomes. That's what they're used to looking at, and so this distinction is a critical one.

The second big lesson that has been learned is, we don't really achieve much in the way of impacts. We can talk all we want about the little asterisks and statistical significance and all that, but actually the vast majority of social programs that have been evaluated, if you really get down to it, do not accomplish much in terms of long-term change in the lives of humans in them. Now when that first started popping up in the first round of evaluations in the 1970s, we looked at it and said, we oughta to try something different. Well, we tried something different in the 1980s, and we tried something different in the 1990s. Now it's a big country and I know there are some evaluations that do show some positive long-term impacts, but there has been a high degree of consistency in the trend in the past thirty-to-thirty-five years, and that has been that it didn't matter what the idea was; it didn't change somebody's life over the long term.

And that's part of the reason I was anxious to write this paper. This is like any field. You try things; you do things. And then after three or four decades, you step back and look and say, wow, that has been the biggest and most important thing that really good evaluations have pointed out. We aren't changing people's lives, on the whole. Why is that? It's a point, for philanthropy, for government, and for those of us who are interested in these issues, that needs to be addressed. Why?

It is not the natural instinct to deal with that issue. What I'm going to say is obvious. It's not government that can deal with the issue because we have a democracy, which means that we have people who need to get elected or reelected all the time. What they have to do is bring up new ideas. And we constantly are getting new philanthropists, now. They're going to have new ideas. So we are not in a field where the incentives or the structure of things leads toward reflection on what these thirty years have meant. We're in a field where everybody comes in and thinks, yes, okay, that didn't work so well, so now I've got something new—right? And that tendency to create something new is never going to go away. None of us need to worry about squashing that tendency. But it leaves little room for the reflection on *why*. Why are we in that position? I'm going to set forth several reasons for this, some of which I think are easier to deal with than others.

One, for going on forty years, now, everybody who's in favor of social programs has vastly overpromised what they could achieve—in part because they didn't think they could get political support without saying that, and in part they say it because they really wanted to believe. The recent example of the after-school programs is a good one. So you just have massive overpromising rather than trying to think about, if you're going to go into a poor community and let's just say you're going to deal with the youth issue, what over the long term would be the necessary infrastructure of things needed for a child to come out the other end and maybe have a

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

significant impact on their long-term life. What is doable and what we can do? Maybe they need good after-school programs plus good summer programs. We need to work on the schools. We need to reduce credit-card—what are a doable group of things?

We have treated each program as if it was going to accomplish the whole thing. It was just naïve. And as I said, it was also out of nervousness that we wouldn't get any support without overpromising. A great example that I participated in was the Summer Training and Education Program (STEP). The STEP program was a summer program set up by Ford, the Department of Labor, and the Department of Education in the 1970s. It was to deal with thirteen-to-fifteen-year-old kids who were already two or more grades behind; in two consecutive summers you would attempt to bring them up to grade and prevent further summer learning loss—which the evaluation showed it did. But the key part of the evaluation and what everybody promised was that five years down the road you were going to have higher graduation rates and less teen pregnancy. And it didn't do that. So the program, which by then was in a hundred communities, was abandoned because it didn't solve the long-term problems.

It's a tough thing in our country to develop, but we have to have the kind of patience to think about infrastructure for people, rather than going for the one thing that's going to solve all the problems. In the past, we've just asked too much.

The second reason, and it's often a difficult issue to discuss, is that the not-for-profit sector is a weak sector. It just is. It's not that good at doing what it does. And in part, that is just the structure of things. In every other industry in this country, particularly in the for-profit sector, when you invent something new, whether it be an airplane or a car, there are immediately hundreds of companies doing it; a decade later, the market sorts out the 375 that ought to die and the ones that are left; and then they go on and we worry that they are being monopolistic and all. Actually, they were the ones that proved they could do it the best. The not-for-profit world—and it's nobody's fault—does not have a structure like that. And we keep inventing new organizations, and as new philanthropists come along they invent other ones. And we just have an enormous number of organizations out there which are inadequately equipped in staff or infrastructure to do a really good job in what they claim to do.

Not only has that not changed, but it has grown worse over the thirty-some-odd years I have been in this field. During my last few years at P/PV, some of the California foundations asked us to look at fatherhood programs in the San Francisco area. San Francisco is a major city but it is not a huge city by population standards. We found 272 not-for-profits that called themselves "fatherhood programs." And when you probed just a little bit at what they did, they were very weak programs. They often had very charismatic, committed, heartfelt people leading them. Underneath it, it was thin at best.

And what I worry about is that when programs are evaluated—and some evaluation firms have learned to do this better—there's nobody out there ensuring whether the program being evaluated has some quality to it. And when you set up a big demonstration—I know because I participated in doing this—you go out sometimes and you either set up organizations or take organizations that have never done such work, and you get them to implement some new idea.

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

And you do it because you give them plenty of money, because you're in a hurry to get the numbers so that you can impose an evaluation on it. And then five years later you get the results, and if you read beneath them, the results are weak. There's no question of that. What the results also show is that they did a lousy job. They really didn't to a good job of delivering any of these services.

Now, many people had hopes—Bill (Schambra) and I talked about this—that with this new infusion into philanthropy of people coming from the business sector some of this would change. I'll just tell you bluntly, I see no evidence of that whatsoever. And I think if you talk to a lot of people even in the foundations that have been set up, privately you will hear it, too—because the natural tendency of somebody who has made billions of dollars in the private sector is to come into this sector heart-first. They just don't want to do all that tough stuff they've done before. They don't know this sector. So they make call for evaluations and all, but what they're not looking at is the quality of what's going on in the service.

I'll just give you one quick example, and I won't use any names, but there is a very well-known philanthropist who made billions of dollars in high tech (laughter) and who came in with the best of intentions to really do it right. They were going to pick programs slowly and they were going to bring in a high-priced business consultant and get a strategic vision. And when they had done all that, they called me down and asked P/PV to do the evaluation. And I had a 150-page strategic vision plan and the director of this program was really a charismatic, smart, committed person—no question of that. And I said, sure, we'd love to do the evaluation. The only thing is, we need a month to look at the program. And I sent my staff down to look at the program—it was in literacy development—and they told me that it was off-the-charts bad. This very smart man who had made billions of dollars had never had anybody sit down, talk to the staff, watch the kids, see who came—in short, see if it was any good. And to me, that factor is still missing amongst donors. Certainly the public sector is not good at that issue. You would hope philanthropy would be but they're not.

So, I would just say this. I'm not against evaluation or measurement at all. It's just that I think we are at a point in the history of policy and programming that we've built an industry that knows how to do all this highly technical, methodological stuff well with regard to finding impacts. That's great. But the issues that we need to focus on and focus philanthropic resources on are these: What does a *quality program* actually mean in literacy and after-school programs? That is not unknowable. It might have seemed unknowable forty years ago that you could do random assignment in social programs, but it *was* knowable. Just the same, today you can find it out for literacy programs, you can find it out for certain after-school—you can develop metrics around the issue of quality. How funders can help programs achieve quality, how the evaluation industry can be turned toward looking at quality, is by asking, what are the metrics of actually what you do on a day to day basis and is it working or not? Not five years down the road. That is the turn that we need to take in how resources in evaluation are spent.

I also think that at this point in the history of social programs, we need to address the really big issue, the infrastructure issue. I think part of the reason hasn't been dealt with well is because the conservatives and the liberals each have a piece of the truth and the foundations and particular

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

administrations can never get them both. A liberal foundation doing what they think is needed so that kids in a low-income community stand a better chance will go into that community and give a lot of money to social services, and they'll want to change the way the agencies work and all. A conservative foundation will tell you that to solve the same problem we ought to be concerned with getting crime off the streets and better parenting. And they're both right, you know?

Until we can get to the point where we can deal with that issue and see each of these programs as fitting in that infrastructure; until we can make sure that they're all high quality and thin out and not fund the ones that aren't; I think in the next thirty years we're going to see this same revolving door of, we'll have some new ideas, we'll have some elegant evaluations, and we'll have weak results just as we have had for the last thirty-five years.

(Applause.)

KATHLEEN ENRIGHT: Thank you, Gary! How many folks were able to read Gary's paper in advance? It was really a great piece and it was so good to hear someone with your evaluation credentials talking about capacity building—basically, you're talking about the fact that execution trumps strategy sometimes, and that's something that I wholeheartedly believe.

Before I get going, I have to say that I'd like to take exception with Bill (Schambra)'s characterization that whoever wins the NCAA tournament is the best team. I am a Kentucky fan; I grew up in Lexington, Kentucky. How many real diehard basketball fans are here today?

(Crosstalk.)

Okay, so five of you will get this analogy. If you saw the 1992 game between the Kentucky Wildcats and Duke, you cannot proclaim that the best team ultimately ends up winning the tournament. That was luck and some, you know, chicanery and maybe bad refereeing. (Laughter) But anyway, very passionate—Kentucky lost, in case anyone was wondering about that. (Laughter)

AUDIENCE MEMBER: Do you hold a grudge?

KATHLEEN ENRIGHT: I do, I do—1992, God love it.

Well, I'm Kathleen Enright, as Bill mentioned before. The fact that I run a group called Grantmakers for Effective Organizations (GEO) would probably make you suspect that I, myself, am a metrics maniac. It turns out I'm maybe not so much a fan of metrics, and so it's probably going to be as surprising to you as it was to Gary when he and I talked this morning about our perspectives on these things. GEO, for those of you who aren't familiar with us, has very deep roots in both capacity-building and evaluation. We are the community of grantmakers that focuses on performance improvement, both philanthropic and nonprofit. But our particular approach is that there are a whole lot of things inside grantmakers' control that, if they were to make some minor adjustments, would ultimately help the nonprofits they support achieve a

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

whole heck of a lot more. So our mantra is that grantmakers only succeed when the folks they fund achieve something meaningful. So we're not the ones doing the work.

That has a lot of implications for how we think about evaluation. We now have some of the thought-leaders on evaluation within our community, and over the years the community as a whole has gone through its own revolution in the way that we think about evaluation. And my sense is that there are four fundamental shifts we need to make in order for evaluation to really yield what we hope that it will yield.

The first one is about proof. Many times, when we think about evaluation, we're looking for proof. Did that thing work? But instead I would suggest that we need to think about evaluation as a mechanism for *improving*. So rather than thinking about proof, we should be thinking about how can we improve, how can we do this work a little bit better the next time. And there's a good historical basis for why we were looking for proof. In the years past, philanthropy was the social innovators who would test a program, and if the program was found to work, the federal government would pick it up—for example, Head Start, and there are others that went that way. That's not the era we live in anymore. So we're doing evaluations that are no longer for the moment in time that we currently inhabit.

Also, on that “proof” point, how many of you have heard either a board member or a funder or someone say, “I just really want to know what the SROI is. What's the social return on investment of your work?” In other words, they want to monetize it. They want to get it down to a number. Well, at a fantastic organization called REDF (formerly known as the Roberts Enterprise Development Fund) out in the Bay Area, a guy named Jed Emerson actually set about to find the social return on investment in a particular subcategory of nonprofit work, getting really hard-to-employ people into jobs. And he did it! He did it. It took a really long time, a huge amount of money, and enormous amounts of effort on both the side of the nonprofits and the foundation, but in the end they were able to say that certain job training programs relieve  $x$  number of dollars off of the social safety net because the people enrolled in them are in sheltered workshop jobs at first, and then they are able to reenter the mainstream economy. So there's a number.

Well, there are a couple of things wrong with getting to that number. The first is that it didn't actually drive resources. These organizations *had proof* that their work actually was making a difference and was relieving the social safety net programs of a burden. But dollars didn't follow.

Second, it didn't give them any sense of what they could do better in order to achieve more. They didn't know, for example, that if they just made a little shift, if they handled these clients in a slightly different way or if they provided this additional support, they might be able to achieve more.

REDF no longer uses that analysis that was so hard-fought to win.

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

I believe, and what GEO has come to believe, is that the highest calling for evaluation is as performance improvement. These are complex, sophisticated issues we're working on, and it's nearly impossible if not completely impossible to get to a number or to get to definitive proof. I've got a lawyer on my board who says that evaluation is like the preponderance of the evidence. If the preponderance of the evidence suggests that this is making a difference, then you've got to be satisfied with that. We have to help educate our boards to understand that that's all we can realistically expect to understand.

And it's really based on the fact that social change or progress doesn't happen in a predictable causal way. That's one of the things that gets gummed up at times in evaluation. You use a tool that's really good for one kind of problem to try to figure out something about a very different kind of problem. There's a very smart evaluator named Michael Quinn Patton who describes it this way. He talks about the fact that some problems are simple; some are complicated; and some are complex. A simple problem is baking a cake. You can use a logic model to figure out whether or not you added the right ingredients, you baked it for the right amount of time, and whether or not the cake is going to come out correctly. Then there are complicated problems, like sending a rocket to the moon. Yes, it's incredibly difficult to do, but if you stick to it and you do the technical fixes that you think you need to do, you can get there. And then there are complex problems, like raising a child. Now if you input really good daycare at first and then the best possible preschool and three square meals a day and then really good colleges, you have no guarantee that the output is going to be a well-adjusted child. There are too many externalities. There are too many things that are outside of your control. Any parent here can attest to that fact.

So we're using tools that are most appropriate for simple problems to try to address highly complex issues. In sum, the first of the four shifts I mentioned above is getting off the bandwagon of trying to *prove* and focusing more on *improvement*.

The second shift for philanthropy is to stop worrying about attribution—what did my money buy?—and be satisfied with *contribution*. The same points apply there. It's just not the best way to spend dollars to try to determine causality. As most of us in this room know, the average grant size at a foundation is \$50,000. So it's just completely unreasonable to expect that with \$50,000 you're going to be able to attribute anything. It's just sort of idiosyncratic and not completely helpful.

The third shift, one that I think is absolutely imperative if we're going to get the highest value out of evaluation, is to stop doing top-down accountability and to think about *mutual* accountability. How many folks are in philanthropy in this room? Not many, right? (Show of hands.) You're the subjects of them? Good, there's a good percentage. And how many are the subjects of philanthropy's push for metrics? (Show of hands.)

You know, evaluation is kind of crummy because it feels like something that's being done *to* you. Someone's coming to evaluate you. If we rethink the way that these problems are going to be solved and realize that no one organization is going to be able to do this on their own; that the problems are entirely complex; and that our success in philanthropy is completely dependent on

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

the folks that we fund, we've got to get away from top-down accountability, have some skin in the game, and hold *ourselves* accountable for the performance of our grantees.

There are some good ways that that's happening. The Center for Effective Philanthropy has a grantee perception tool with which you can tie grantee feedback to individual program officer performance. So you can use that as a way to evaluate individuals. But we're pretty far off on this one. We're not anywhere close. And one piece of evidence of that is from a study that GEO did a couple of years ago in combination with the Urban Institute. It was a survey of every staffed foundation in the U.S. And there's this sneaky little methodology whereby on the front end you ask, "What do you think is important to being effective?" and on the back end you ask, "What do you do?" One of the questions was also whether in the past two years the foundation had solicited feedback from its grantees. Nineteen percent had done that. Then, somewhere else in the survey, the participant was asked to rate overall grantee relationships—are they fair, poor, good or excellent? Well, of those grantmakers who said they had good or excellent grantee relations, only twenty-nine percent had asked for any kind of grantee feedback in the previous two years. So we are certainly *not* being mutually accountable; we're not holding ourselves up to the same lens, the same set of standards as we expect of our grantees.

The final shift we need to make is that we've got to move away from metrics mania to assessment that is developmentally and organizationally appropriate. Again, with the \$50,000 average grant size, let the punishment fit the crime. How many of you have filled out fifteen-page reports on a thousand-dollar grant? I have. And I have to say, one of the tough things for us is that I now feel an obligation to give that feedback to the grantmakers that give *us* grants. I've had some interesting conversations there. And we have to be okay with the fact that we might not get funded again, because otherwise I would not be faithful to GEO's mission if I did not give that kind of feedback.

So there is no one-size-fits-all. And there is some hope. Many foundations are rethinking the way that they ask for metrics and evaluation. And there are really compelling reasons for us to do this. One of them is that the Daring to Lead study by CompassPoint and the Meyer Foundation a couple of years ago found that we in philanthropy are contributing to executive burnout. That was one of the top contributors to executive burnout. Many executive directors spend eighty percent of their time on the application and reporting around fundraising. That means they're only spending twenty percent of their time on programs. Think about if that were shifted. And philanthropy has to take the responsibility for that. That's a compelling reason for change.

Part of GEO's advocacy agenda is to give people permission *not* to evaluate sometimes, where it's inappropriate to do so, and to ask just the right questions about what success will look like and how you'll know that you got there, rather than always mandatorily going through a drawn-out process to get to the evaluation results. So I was absolutely thrilled to hear someone who is so well-regarded in the evaluation field talk about focusing on execution, focusing on quality as a precursor to thinking about how you might want to look at impact or outcomes. And I look forward to the conversation with my colleagues here in a minute.

Thank you. (Applause.)

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

HOWARD ROLSTON: Good afternoon. It's a pleasure to be here. It was a pleasure to read Gary (Walker)'s paper. There is very much that I agree with on it and there are some places where I don't entirely agree. Maybe it goes back to 1976, which is actually a little before I got into this area—not because I wasn't old enough, but because I wasn't really in it. Gary was in the part of supported work that dealt with the unsuccessful parts, and the part that I ended up being with was the one success in supported work, which was the welfare to work side of it.

It's so nice that Gary (Walker) put these lessons down so clearly in his paper. It makes it very easy to think about these things. And he is exactly right; all I can say is yes, yes, yes. Gary's lesson number one, I think, is that if we have learned one thing, we have learned that outcomes are not impacts, and that outcomes can be a very misleading measure when we do know what impacts are. What that leads me to think is not that we can do without evaluation or that evaluation doesn't have to be the centerpiece of how we can improve programs, but rather that we need to be very strategic about how we go about using the tools that we have. And I agree very much with Kathleen (Enright) in saying that it is absurd when people say, I'm going to hold somebody accountable for the impact of a \$50,000 grant. I think it's pretty silly even with very much larger grants than that, given the resources it takes to do things right.

But I think in the end how we will make progress in social policy and social programming is by learning that what makes a difference—in terms of what this program is doing, what this reported improvement in programming is doing, or whether something is a quality program or not—ultimately comes down to whether the outcomes are better than they would have been. And I think we have to face up to that, and to the fact that there are important measures of quality—like if nobody shows up or there is an instructor and the instructor doesn't show up. There is no question that there are sort of broad operational measures that make sense and that we need to know about and that we should certainly be trying to improve. But in the end the real measure of quality is, did it make a difference? Were the outcomes better for the people who participated than those who didn't? That requires evaluation. It doesn't require performance measurement. It requires establishing something better than what otherwise would have happened for people.

Now, the place where I think I differ the most from Gary—although I'm not sure how much it is at one level or another—is that I'm much more optimistic about evaluation being able to deliver and the fact that social programs at least in some areas have made a difference. I'm going to talk today about the area with which I'm most familiar, partly because I'm most familiar with it, but primarily because it's the area where there has been the most consistent application, over a period now of forty years, of randomized trials—and that's the welfare-to-work area. I'm going to talk about the results, just to give an illustration of how I think the results are better.

But it also gets us more to the question of expectation than Gary's summary. I'm going to talk about a meta-analysis. What was in this meta-analysis of mandatory welfare-to-work programs are about a hundred estimates—that is, a hundred different programs that were evaluated using experimental methods. We are going to talk about their effects on people's improved earnings. I think it's important to realize that we are talking about very large numbers and that makes a

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

difference; we are not talking about three over here in one social program, and four over there in another, and I think that is part of the lesson.

Now the summary is, the average program—not the best, but the average program—improved people’s earnings over a period up to a peak at three years. At six years there was no difference between the experimental group and the control group. That’s for the average program. Now that is evidence, it seems to me, that at least in this area there is some progress. You can debate about whether six years of earnings is enough or not. And I suppose we would have liked it if the experimental group’s earnings had just kept growing, that they kept beating the control group. But I think in most areas we don’t expect that; for example, that is not how we generally look at many health treatments. If you look at medicine or things like that, except in the case of things like vaccines or the rare kind of cure or prevention, we look for a modest difference over a reasonable period of time. Now I agree that for a children’s program, say for a program for preschoolers, if it improves their reading, their pre-reading skills for a year and then it disappears thereafter, that is not of any great significance. But if a family is more dependent on earnings and less dependent on welfare for six years, even if the control group eventually catches up, I think that is an achievement, and it’s not one that we should feel is not worthwhile. So I think there is evidence that as a matter of fact we can produce systematically better outcomes. The question really is, can we then by a program of continuous improvement and evaluation improve by another ten percentage points, and then try to beat that standard?

So I’m very much in line with what Gary (Walker) says about raised expectations. We’ve sort of set up these programs, that if they don’t make a lifetime difference in the end we all end up with the same outcome. The question is, can we have some sustained, serious improvement, and then can we build on that? If you think about what it is we are spending, for example, if we’re now spending \$1 trillion on cancer research, are we doing better than we did thirty years ago? I think the answer is yes, but it’s not because we’ve found a cure. I think if we think about social programming in that way, well, we haven’t cured the problems. It’s totally unrealistic to think that we can cure problems *that* quickly, and it’s totally the wrong expectation to think that we can learn about how to do it that quickly. But at the same time, if we don’t undertake these kinds of serious evaluations and have a serious way to do that, then what reason is there to believe that we are going to learn at all?

So I agree with Gary’s call for patience. It’s not that we can sort of give up evaluation for other measures of quality, although I think it’s important and I think Gary’s right when he says that we do a better job of looking for quality when we invest in a certain kind of evaluation than we did, and we can do even better.

But I think that the main thing we have to think about is how to be very strategic. If we think about a particular area of social programming—and we have to think about them separately; there is no sort of thinking about social programming in general—we have to focus on answering a few questions well. What could we learn about this particular area of programming that might be improving outcomes? How do we build an agenda so it’s not just a single experiment that is going to answer the question? How can we build a series of evaluations so that ten years from now we will know more than we did, and learn from failures, too, if we go about it

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

systematically? One of the great problems is that we don't learn from failure because we haven't set things up systematically enough to create the agenda in the first place. But if we have an agenda and we can make progress incrementally, then I think we are doing something.

So we need patience. We need reasonable expectations. Funders sort of give up on the idea of managing to outcomes instead of managing to design. If we look at a particular area, the question is, what do we know about this area, about improving outcomes? If we don't know anything then we ought to try to start learning. But if we know something—that it's getting people there or designing the program in a certain kind of way—then *that* is what they ought to hold their grantees responsible for, for doing it that way. If you think about it, when we discover a medicine that does better than another one, then we prescribe that medicine. We try to make sure that people take it regularly; that that is what doctors prescribe; and that people actually take the medication that way. We don't—at least if we're not foolish—say that were going to hold you accountable for the outcome. No, we are going to hold you accountable for doing what we have evidence for works. And in the meantime we are going to try to figure out what works better. It seems to me *that's* the model we need—not a model that says, we gave you a dollar and now we're going to make sure, as Kathleen (Enright) said, that that dollar produces something.

Thank you. (Applause.)

ALBERT RUESGA: Thank you very much. I also want to thank our hosts Bill Schambra and Krista (Shaffer), and do a big shout-out to my colleague Kathleen (Enright), whose organization, GEO, I think is one of the few in philanthropy that speaks with any real wisdom about the subject of evaluation.

Bill saw my role on this panel as the guy who when asked about metrics replies “metrics schmetrics,” but I'm going to have to disappoint him—well, just a little bit anyway. Measurement and evaluation, when done properly, are not just a bit of value added to a philanthropic or nonprofit piece of work; they're absolutely essential, and only a fool would disagree with that proposition. Here I mean not just the kinds of formal evaluations described by Gary Walker in his essay, but informal evaluation as well, the kinds of course corrections that all of us naturally make when we embark on a project, take a false step, and adjust what we do accordingly. Evaluation is not and should not be the sole province of highly compensated consultants. We evaluate all the time. Our eyes and ears notice things the most astute consultant will never notice. And we will often be our own worst critics.

Now here is where the “metrics schmetrics” comes in, perhaps. More nonsense has been spoken and written about evaluation than about just about any other topic in philanthropy. The number of people practicing evaluation without a license and without proper grounding in the scientific and philosophical intricacies of the subject is, in my view, a scandal. And worries about evaluation engendered in part by logic models the length of whale intestines have become the math anxiety of the philanthropic world.

And my general thesis, if I could call it that, is that from the perspective from somebody like Gary, whose organization has been commissioned to conduct lucrative large-scale evaluations of

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

social programs—I'm talking about "lucrative" in the nonprofit world, now, the "impact revolution" might seem like a good thing. But from the ground, from the perspective of many people working in community-based organizations on the front lines of these very intractable social problems, this so-called revolution has brought with it new sources of irritation and new ways of adding meaningless make-work to already overburdened nonprofit staff members. It has not been, in other words, a "people's revolution" (laughter), but rather one championed by elites, like myself I'm afraid, unable to see far enough beyond our own measuring sticks to understand what the limitations of formal evaluation are, and the trade-offs in staff time and other resources that these formal techniques require.

Now some of these limitations have been well rehearsed; I'm not going to go into them here. I'm thinking, for example, of the charge that this revolution has in many cases—not all—prompted us to measure the things that can be measured rather than the things that are important. Other limitations like the absurdity of most logic models are more technical and less well understood and discussed. And I'll get to what I believe is perhaps the most disastrous aspect of this revolution in a minute, but first I'd like to make a few quick comments about Gary's paper.

After pointing out that outcomes are not the same as impacts—and more on this distinction a bit later—Gary comes to the startling conclusion that quote "the overwhelming majority of social programs with impact studies do not show a significant change in participants' lives a year or two after the program." He tries to soften the apparent harshness of this conclusion by claiming that these dismal results are an artifact of our approach to conducting these impact studies. He points out, for example, that these studies are typically conducted for new projects that haven't been given the time or resources to work out their kinks. So you have new programs, lots of problems to work out, dismal results.

Now, as you ponder Gary's claim, I urge you to keep the following in mind. Those of you, like myself, who have worked at federally funded programs like those evaluated by Gary's organization know how notoriously stingy the feds are when it comes to payments for program expenses and overhead. Throw in the byzantine reporting requirements and I'm surprised that any of these programs succeed. And it's difficult to assess Gary's claim without having all the data in front of us. Is it possible that the same organizations that were conducting these impact studies would then be called on to fix up the mess, in which case, couldn't this have introduced a source of bias? And given that Gary himself says that impacts might not be apparent for ten or fifteen years, how many of the impact studies he refers to are conducted over this very long time span? How can any impact study, *any* impact study, with a pretense to being scientific ever control fully for the bias introduced by the self-selection of the clients for a particular program? And the list of methodological worries goes on.

Unfortunately, the ink has not had time to dry on Gary's essay before his words are yanked out of their context and given pride of place in the invitation to today's panel discussion. And here I quote, "Why is it that philanthropy has learned so much about metrics and yet has so little by way of measureable success to show for it?" And I have to vigorously reject both of these implied claims. Number one, I'm not convinced we've learned much about metrics. We're doing more of it, perhaps. I'm not convinced that we're doing it better. And I certainly reject the notion

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

that philanthropy has little to show by way of measurable success—but the damage is done, Bill (Schambra), right? The mean has been loosed on the world and there's no way that I'm going to call it back. (Laughter.)

GARY WALKER: You're doing a good job, though!

ALBERT RUESGA: Oh, thank you. I resonate most with Gary's essay when he writes, "The first things funders need to be accountable for is the *quality* of the program which they're funding. That requires patience, and a use of funds for things like training." I can think of other investments that funders should be willing to make, but I like very much how Gary's words suggest a certain amount of care and support for grantees, and camaraderie and collegiality.

And this gets to my primary beef, really, with this whole "metrics revolution," the disaster I mentioned earlier. I find the image of a funder with a stopwatch in one hand and a clipboard in the other, hunched over a perspiring grantee, rather ghastly, frankly. (Laughter.) It's uncivilized. It's so clearly opposed to what I believe should be the ethos of the charitable sector, an ethos rooted in love for our fellow men and women, expressed through our work, and incorporating of the values of cooperation and mutual support among many other values. And too often, unfortunately, I see funders using evaluations like blunt weapons, barely understanding—if at all—the limitations of these tools and certainly being unwilling for the most part to turn these weapons on themselves.

I have a very few quick points about the impact revolution as seen from the ground, and then I'm done, I'm over and out. I'd like to suggest something called an "outputs and outputs counterrevolution" to the impacts revolution. I find the whole progression from outputs to outcomes to impacts one of the great bugbears of contemporary thinking about evaluation. And for those of you unfamiliar with this bit of chicanery, outputs are the things that you do; outcomes are the things that happen because of the things you do; and impacts I assume are outcomes that stick. We can think of them that way. So, say the thing that you do, the output, is mentoring a child five hours a week. There are some kinds of outcomes for this child, and the impacts—I can only guess at what those impacts might be. They grow up to be President of the United States, or whatever they might be.

The next thing that we're going to be requiring of grantees beyond impacts is—I don't even know what the word for it would be—maybe hyper-impacts (laughter), and this would be the effects of their programs on the next life (laughter), or on universes parallel to our own. I want you to note that it would be absurd for us to call the gas company, thank them for their outputs, *their outputs*, namely the gas that they deliver to our homes, and then complain that they haven't demonstrated to us any outcomes or impacts. And why is it, and I leave this as an exercise to the audience, why is it that we reserve this nonsense for the people who work in the nonprofit sector?

And what is it that's so bad about outputs? As a donor, it's enough for me to know that you've delivered a quality youth development program to twenty-five kids in a church basement who would not have had the opportunity otherwise. And for God's sake, don't incur the expense of

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

trying to track the effect of your program on these kids ten or fifteen years down the pike; that would be a ridiculous waste of resources. Unfortunately, all this talk about outputs, outcomes, and impacts blinds us to the fact that in many cases—again not all—simple outputs are all that we can reasonably hope for or require.

I don't know how many of you have followed the development of distributed computing. In this model, instead of having and running an entire application like Photoshop on your desktop, it's distributed across two or more computers that are connected by a network. There are many advantages to this model, among them your ability to access and use the most updated version of Photoshop, or whatever application it might be, without having to purchase the entire thing directly and load it onto your machine. And just like we've seen the advent of distributed computing in the digital world, I'd like to suggest we try something like distributed evaluation in the nonprofit world.

Here's how it would work in the case of a youth development program, to take a typical example. We assume that academics and others have already researched to death and determined those elements of a youth development program that are likely to yield good outcomes for young people. If you're a grantee, suppose we make it your responsibility to demonstrate that you've incorporated these elements into your program? We then make it part of my responsibility as a funder to know, because I've done the homework and read the literature, what these success generating characteristics are, and to verify that they're characteristics of your program. This is very much like what Howard (Rolston) is suggesting. In this way the burden of evaluation is shared three ways, and neither the funder nor the grantee has to prove for the eleven billionth time that young people respond well to nurturing environments that stimulate their hearts and minds.

I want to make it clear that I'm not at all anti-evaluation; I am concerned that we tend to seek a kind of scientific or a moral certainty from a formal evaluation that it can never provide, and the questions that funders most often bring to an evaluator—was this program worth our \$25,000 investment, and should we continue funding it?—are questions that only the donor can answer. There's simply no absolute scale against which an evaluator can measure the value of some charitable program. Say we measure a 25 percent drop in the truancy rate for one hundred kids in some program and a 25 percent increase in their test scores. Is that worth \$25,000 to you? Each donor needs to answer that question for him or herself. And as donors we will never be absolved of our responsibility to use our good judgment in these cases; and an evaluator cannot give us this answer.

And finally, one of the great benefits of an organization like GEO, one of the great benefits it can provide to our field, is not a training on how to conduct evaluations—I think we have plenty of those—but rather, a training on the questions that evaluations will never be able to answer. We might also benefit from being reminded that in a business context we often strive to convert all our currencies to a single coin, namely money, but that in many nonprofit contexts, values like mercy, justice, love, frequently motivate decisions that don't always make sense to the bottom line and whose effects can't always be measured. So, yes, Bill (Schambra), “metrics schmetrics” to those purveyors of evaluation, the majority in my view—and present company excluded, of

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

course—who: a) don't really know what they're doing, and b) don't really know that they don't really know what they're doing.

Thank you. (Applause.)

WILLIAM SCHAMBRA: Well, Gary, any thoughts in response to what you've heard?

GARY WALKER: You know, it's complicated. I wish I could be really simple-minded about this. Everybody had really good contributions here, and in some ways it didn't sound like there was an awful lot of disagreement. Even at the end, Albert, you put it really dramatically and my heart agrees with a lot what you were saying.

But what I don't believe is that we're going to do well in this world of social policy if we don't have some way to measure that something good has happened. The politics of this country, to my mind, are sort of always fundamentally centrist-conservative, although there are little blips in the '30s and the '60s. Where I agree is that we do need to push more towards a mutual responsibility that the programs themselves are quality. I don't think that I have the confidence maybe that others have that a very high percentage of programs are very quality programs, or that they deliver that much for kids. I don't think a good heart is enough in today's world to run a good social program.

WILLIAM SCHAMBRA: Is it your impression, Gary, that we know enough, really measurably enough about what works in a variety of human sectors, that there is a whole raft of practices that we have available to us that we are somehow overlooking? I am just trying to get a sense of *your* sense of where we are in terms of an inventory of practices that are useful and that could be applied. One thing I have heard from all of you, it seems to me, is that we should apply what works, that we *know* what works. There is some disagreement about how you go about figuring out, or how important it is that we go about figuring out, what works, but the consensus seems to be that we should apply what works.

And before you answer that—or as you answer that—do reflect for a moment on the fact that philanthropy, as your paper points out, is notoriously allergic to doing again what has already worked so well. Philanthropy is about experimenting, the cutting edge, and a new and untried program; that's the great virtue of philanthropy—not applying what we know, but being at the frontier, experimenting and trying to figure out answers to what we don't know.

Do any of you on the panel want to take a stab at that?

HOWARD ROLSTON: I think that the answer is, by and large we don't know that.

WILLIAM SCHAMBRA: We don't have a large—

HOWARD ROLSTON: I think there are areas where we do. I could point to studies and synthesis in the welfare-to-work area that pretty conclusively show that if you do it this way you will do it better than if you do it this way, or you will get different impacts. I also think that at some point we might talk about impacts; I think Albert was using the term impact entirely

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

differently than Gary and I were using it. I think that for clarity we might get to that. But I think the answer is, no. We see what it takes to have that; we haven't spent enough time at it, and we haven't spent enough resources in figuring it out in a systematic enough way.

KATHLEEN ENRIGHT: I'll add to that. I think it varies broadly, sector by sector. In certain sectors we've made a lot more progress than in other sectors, and oddly enough the ones where we've made more progress are the ones where things are a little bit more knowable. You can figure out in community and economic development what kinds of things might help turn an urban area around. And so there are areas that are much more sophisticated. But even the "stuff" that's known is not at all well communicated. There is a huge disconnect between doing the exhaustive study, printing up the four-hundred-page report, and getting it in the hands of anybody who can do anything with it. And so there is a nice marriage between the evaluation field and communicators that hasn't yet happened and really does need to happen.

GARY WALKER: And I would say, no, we don't know in all areas what to do. But oddly enough, we know a few generalities that actually apply to a broad number of programs. We know, and I think we know, both through evidence and after a while just common sense, that in a lot of communities, for people who for whatever reason are not in the mainstream, who are behind and are young, it is going to require a consistent, semi-long-term human contact to keep them motivated to do well. There are these programs where you go in and you do something for six months, and then the people who have participated go back out and face the same temptations, the same world; what really keeps people in line is continuous human contact. Mentoring does some of that; you can do it in other ways. We know through David Olds' program how to go into hospitals and start dealing with mothers very early on and to actually have some long-term impact.

But there the problem is, and I think what Olds has shown is, a program works and it works in different environments *if you do it like he did it*. And now we run into the nonprofit culture, which is one of, every locality is different, everybody has got to have commitment charisma, and none of it wants to be like McDonald's. But you know what? There are some things that it would really be better to package and take more broadly because you would get more impact. And the truth is, in dealing with human beings there is always plenty of room for creativity, but we just don't even allow the packaging of a few of the essential things we know—and that I think is a real problem. A different one than what we have discussed here today, but a real problem.

KATHLEEN ENRIGHT: Yeah, absolutely.

ALBERT RUESGA: Ibid. and op. cit! (Laughter). I agree, there are many areas where—there are many studies done about youth development programs, access to primary care in the health field, community development, development of housing, and ending of homelessness, over and over again. It's not that the knowledge is perfect, but we have a pretty good idea of the kinds of elements that should be present in certain kinds of programs to achieve certain kinds of outcomes and ultimately deeper impacts. We depend on these when we do our own grantmaking, that our program officers become familiar with these. One thing that has been lacking until fairly recently is a kind of central depository for all of the evaluations that have been conducted by various

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

fundors, grantmakers, and other people, of these kinds of social programs. Fortunately, the Foundation Center ([www.fdncenter.org](http://www.fdncenter.org)) has a new resource, PubHub (online at <http://foundationcenter.org/gainknowledge/pubhub/>). It's really a collection or database of these evaluations that have been done by various foundations over the years, and I think the next step is to annotate and have the users of these evaluations actually comment on them as far as their utility for new donors, people who are thinking of moving into a particular area. This has been the answer to a prayer many of us have had about learning from what has gone before rather than trying to invent things anew.

WILLIAM SCHAMBRA: I have yet to see a foundation, however, describing itself by saying, don't bring us your new, innovative ideas; we only fund things that are proven to work; chances are we are only going to fund after you've been through your start-up and after you've got the program running; and, we'll come in and support your efforts, but after the fact. I mean, isn't it rather the case that almost all the foundations say, we fund only new and innovative programs; don't expect us to be around past the start-up point; and don't expect us to fund past the first three to five years of funding? In other words, isn't there a fundamental clash between the notion of cumulative wisdom in the field and one of the fundamental practices of philanthropy, as Gary (Walker) pointed out in his paper? And how do we deal with that, if I may ask?

GARY WALKER: You know, I think that this is an interesting point, and I think that you're certainly right in terms of all of the new people coming into philanthropy. Every time you get a bunch of new people coming in, I do think that's the case. What's interesting is that the form of philanthropy—the kind that exists forever and turns into these big bureaucracies; the donors aren't around or aren't alive; and it's the easiest to criticize—is oddly enough the kind most likely over time to begin to change this practice. Look at the Edna McConnell Clark Foundation, which is doing exactly that; they only want to take things that are out there, nonprofits doing things that are at least apparently working; they want to invest in organizations to grow them bigger and better and stronger. So it's an interesting distinction in the field of philanthropy.

ALBERT RUESGA: I think your claim needs to be evaluated Bill. (Laughter.) Formally.

WILLIAM SCHAMBRA: (Laughter.) Formally, that's right.

Howard (Rolston), you heard Kathleen (Enright)'s description of the kind of more modest expectations for understanding and, if you will, evaluating what these programs are that foundations are supporting. It's a much more modest approach. Is that a way of building knowledge, or are foundations sort of hopelessly trying to talk the language of evaluation and knowledge building without taking the steps, i.e. randomized studies and randomized control groups, to really be able to say with certainty that what they're doing is working?

HOWARD ROLSTON: I don't even know that we learn with certainty through a randomized trail. And it's important to distinguish—the word “impact” means something entirely different in the context of a formal evaluation than in the context of a logic model, where it just means an outcome that is further down the path. If somebody went to a program and we then measure later on what we were hoping to achieve, unless we know that that's somehow better than what would

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

have happened had the person not gone to this program or if we are comparing two different programs, then we don't know whether the program made a difference, and therefore we don't know whether it was a worthwhile investment. Unless there is some systematic way of doing that, we don't learn about how to do things better.

At the same, I entirely agree that that can only take place in a certain setting with substantial resources on a very limited set of circumstances. It would be nice, say for example, if people who were funding youth programs both in government and out would say, okay, this is what we are going to try to learn over the next ten, fifteen years; if they would coordinate their efforts; and if they would try to make some kind of headway that way. But I agree entirely that going out and telling somebody to whom you've given a million dollar grant that they should now take ten percent of it and figure out what impact it had is not going to work.

KATHLEEN ENRIGHT: That's not a lot of money.

HOWARD ROLSTON: Spend the money on programming and hope it's working, and in the meantime take some other money and try to build knowledge.

WILLIAM SCHAMBRA: As you know, this notion of constructing logic models is longer than whale intestines, as Albert pointed out. This is now absolutely the standard practice. No one dares undertake a program without constructing a logic model, but one that is outside this realm of randomized control groups. So is it pointless to do that? Is that a foolish exercise?

HOWARD ROLSTON: I would say no. Logic models are very useful in several contexts. One way they are useful is that one has to think systematically about what one is trying to accomplish and how one is going about doing it, and ask whether there are better ways of thinking about it either from evidence or from common sense. It is good to be thoughtful and systematic when you're planning an intervention in a program. It's also valuable to understand the theory of the program when trying to figure out how best to evaluate it. But a logic model by itself, no matter what you do, is not going to tell you whether the program made a difference or not.

GARY WALKER: Just to understand the logic model, I think the whale intestine length is one good test. But the other one, I think they are great if you would just see the second step taken, which I don't see in dealing with a lot of funders. The second step is, are there too many hard things to do in this logic model? Because if there are, then we either have to scale back expectations or you ought to do something else. You see them put up on the board with like three or four—you know, you start the multiplier effect. If you have six really hard things to accomplish in the logic model, you know it's not going to happen. And that second step is the one I see missing so much.

ALBERT RUESGA: I completely agree with that. I think one of the big problems with the logic models that I have seen is that most of them are dead wrong. They are too simple. First of all, they are not falsifiable, to use an old expression. They are far too simple. And the ones that come closer to being accurate are far too complex to use as the basis of an evaluation. They provide

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

insights; they help organize your thinking; and they provide other psychological benefits; but really there is very little that's logical, in my view, about a logic model.

WILLIAM SCHAMBRA: Good! Let's go to our audience.

STEVE TELES, University of Maryland School of Public Policy: One thing I think that is coming into Gary's talk is, when you think about what your evaluating, when you're evaluating a program, part of that is that we have to decompose what we mean by a "program." A program as a causal theory; we have some treatment and we expect it to have some effect. But we also have an organization that is doing this treatment. So if you get to the end and we have a positive effect, we don't know whether it's the treatment or the organization. It may be that there is a kind of iffy theory, but it's this incredible organization, and when they actually were doing this on the ground it turned out that they were doing something very different than what the logic model they told you was. We know this happens, and in fact, that is what a good organization does. A good organization starts with a causal theory, and then when it doesn't turn out to be doing what they think, they improvise. They are not really supposed to improvise if you're testing something.

One of the implications for philanthropists, it seems to me, is that you shouldn't think of yourselves as testing theories; you should think of yourselves as testing organizations. But the problem is, of course, that the way you test an organization, the way you test their capacity to implement, is very different than the kinds of ways you would test a theory to see whether a theory is going to be applicable over time. And the questions that you have about replication are different, right? People usually think that they have to have a theory that they can replicate at other sites. What they ought to be thinking about is, do they have an organization that they can replicate in other cases, essentially that they can invest in to grow?

So one implication for philanthropy is that you ought to do less; you ought to actually invest in fewer organizations, invest larger amounts of money, and invest over longer periods of time. You ought to make a durable commitment to them, where you're saying, Look, we don't know exactly what's going to work. We know that building up an organization is an organic thing, and that it takes a long time.

It's also the case that even then most organizations don't succeed, so you also have to be willing to pull the plug. That is the only problem that I have with the part about love and caring and all that. At some point you have to be able get your pistol and take the sick dog out.

GARY WALKER: Whoa! (Laughter.)

STEVE TELES: But, again, philanthropists ought to do fewer grants; they ought to have a longer commitment; and there ought to be more of an insistence that what you're doing is investing in an organization and getting it up to scale rather than testing a theory that then can mysteriously somehow get replicated in other contexts.

KATHLEEN ENRIGHT: That's a brilliant comment, and it touches up against one of the main challenges of this metrics mania, which is that there is a whole host of unintended consequences.

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

One of the unintended consequences is this focus on a program and programmatic outcomes with a complete lack of caring about the organization in which that program sits. GEO is an avid advocate for longer-term support and for operating support because nimble well-led organizations are the ones that are going to be able to improvise given the changing circumstances. And that is also one of the downsides of randomized trials; you cannot take a model and plop it down in thirty different locations and have the same effect. There is translatability on some learnings and some key aspects, but you can never rule out all of the externalities to be able to do that.

HOWARD ROLSTON: Yes. Just one thing. I think there is a lot that is true, there, but I think that the main point is: you cannot learn very much from a single site. You can generate hypotheses, and you can say, well, at this point in time this organization with this program produces impact. But that is the jargon that leaves a lot inside the black box. The real issue is the average impact across a set of sites who are the potential implementers of this program. What is the average impact, and how does it vary across sites? If you can't ultimately get to the place where you can do this, so that in effect organizations are averaged over the kinds of places who will be implementing this, then you haven't learned something that is critical to learn. You've just learned that in this place it worked, or it didn't work.

ALBERT RUESGA: I think your comment also gets at another issue that is not captured by many evaluations. That is the issue of quality. And the issue of quality is manifest—I suppose if you think about the outcome of getting enough protein for the day, you might just as well eat at a fine restaurant or Burger King and either way you are going to get your protein. But of course the experience of eating at those two places is going to be very different. And I think that often we overlook the element of quality for social programs because they are serving low income people, and low income people don't deserve to have a good quality experience. You know, this is the kind of thinking where as long as we get the outcomes, as long as they are in jobs, that's fine, but quality be damned. And I think that that is part of what I think is more important for evaluations to capture: what is the lived experience of somebody who goes through some of these programs, whatever the outcomes might be.

GARY WALKER: I thought it was a great comment. I think it really runs against the grain to do what you said, Steve (Teles), even though a few foundations are doing it. And I know you made it for substantive reasons, but I think it is also a smart approach if we can get foundations to invest more in fewer organizations—a smart approach *politically*. It is very funny to look back at the last century and to realize that in the first half of the century, when there was not a lot of federal money in programming, you come out of it with a group of brand-name programs that every citizen in America will recognize: Big Brothers Big Sisters, Boys & Girls Clubs, 4-H, and all of that. For the second half, when enormous amounts of federal money were put in, you cannot come up with an equivalent number of names. This is a brand-name product-orientated culture, not an idea-oriented culture, and the money out there in philanthropy is not spent—and time—paying attention to that political fact either, I think.

EVELYN GANZGLASS, Center for Law and Social Policy: It seems to me we are caught in yet another dilemma. I heard both Gary (Walker) and Howard (Rolston) talking about systems,

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

whether it is community infrastructure, whether it is a continuum, so that zero-to-five intervention needs to be followed up with other interventions, or that there is no silver bullet. You need multiple interventions to make a difference for at-risk youth, for example. Yet we do not measure systems. Nobody is interested in investing in systems. People's eyes glaze over, whether they're in foundations or the federal government. And the federal government and even foundations are set up in very siloed ways, so that HHS evaluates its programs; the Department of Education evaluates its programs; and Department of Labor, its interventions.

I don't have an answer to that; I'm just posing it as a question to you, because it's a measurement issue. It's building political will—because I agree with you, Gary, that unless we measure we are not going to get more investment, and it seems like a syndrome that we are in—we can't measure it, we can't define it, nobody knows what the hell the system is, and we can't get investment in building up that infrastructure. How do we get out of this?

GARY WALKER: I agree with you, Evelyn, and I don't mean to understate how difficult a problem it is. But I think one of the reasons why it is interesting, and I was glad Bill (Schambra) took it on, is, the problem will only start to be solved if philanthropy begins to be interested in it, because of the advantages philanthropy has—they have the independence to take on some of these issues. Consider what happened in California; all of a sudden that state has more after-school money than the rest of the country combined. Even though it's in budget troubles, it has over \$500 million dollars. And the Packard Foundation stepped in and said, we've been ignoring the way the state does things too long; let's put money into how can we help the state do better technical assistance to all of these new after-school programs that are being created. They decided to, in short, put funding in. Well, it was a big thing out there—people said, what?! A foundation is actually helping a state system do a better job?

I think if we start seeing more of that, maybe the idea will catch on. But I do not think the idea is going to come from anybody else. I think the silos in the federal government and state governments are much more in place and impervious to movement than in foundations.”

KATHLEEN ENRIGHT: The good news is, the funders who are out there on the front end of the innovation curve *are* beginning to look at the work from more of a systems perspective. And you can define community however you would like, be it Oakland or early childhood programs or whatever else. But looking at them from a systems perspective—the Monitor Institute, which is the nonprofit work of Monitor Consulting, is doing a sort of demonstration project in Hawaii with a group of grantmakers there. Their project is trying to take up a level the way that think about outcomes and indicators and hold themselves accountable as a collective to some systems-level of change. Barbara Kibbe is working on that, for Monitor.

HOWARD ROLSTON: There is also another way of thinking about getting at the systems issue—which doesn't fully get at it, but potentially gets at a part of it. It is getting a coordination of service-issue, which also gets at the idea that “it takes more than one.” I don't know how many people had a chance to read *The Washington Post* this morning, but in the Metro section there is a report that the Mayor decided to start this plan for kids who are having real problems in school (“New Program To Take Early Action to Help Those Failing,” V. Dion Haynes, March

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

19, 2008, p. B02). The idea is to intervene with the family in a multiplicity of different kinds of ways, substance-abuse counseling and all kinds of things. It all seems like a perfectly reasonable thing to do—to say that the problem is not just in the school; it's potentially in other places and that could potentially require a lot of different services.

Well, they are going to be folding this in in schools. They're not going to go everywhere, but they are going to be spending a lot of money on it, and it's a reasonable idea. So why not do this systematically, and learn as you are doing it whether it's making a difference or not? It's just a shame where there are these opportunities, where somebody is trying to sort of make things work for a family and a child, to pass up the chance to make it an experiment. Maybe we'd know something in a couple of years. As it is, they will do it; they will spend a lot of money on it; and we won't know anything two or more years from now.

DAVID ROODMAN, Center for Global Development: In my mind this institution is associated with the idea that private programs, private philanthropy, is as effective as public—if not more effective. I'm wondering, in this context, if you see evidence for or against that general idea. Do public programs have greater impact, or, moving beyond that question, do public agencies do a better or worse job in general of deploying impact evaluation in useful ways and learning from it? Or more generally, how do they compare and contrast strengths and foibles?

KATHLEEN ENRIGHT: Okay, everyone is looking at me; I'll take this one on. There is just no purity, right? Most public programs are dollars that go either through the state and then to nonprofits, and some of those same nonprofits are also funded by private philanthropy, individual gifts, and a variety of other sources, but I think that public dollars and the kinds of grants that are made through federal and state and local programs are a lot more likely candidates for the kind of evaluation that Howard (Rolston) is so expert in. Private philanthropy just by the nature of its scale is sort of a mismatch in terms of the type of evaluation versus the amount of money and the kinds of organizations that they are reasonably funding. There are couple of exceptions—the Bill & Melinda Gates Foundation, for one. If you have \$64 billion to play with, you might want to be pretty intentional, and they are very intentional, in my opinion, about what is working and what is not.

So, all of that being said—

HOWARD ROLSTON: They sure poured a lot of money into small schools with very little evidence—

KATHLEEN ENRIGHT: And someone got fired for that, you know.

HOWARD ROLSTON: That may be, but we still don't know whether it was a good investment or not.

(Cross talk.)

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

KATHLEEN ENRIGHT: No, no. They have come out fairly publicly, saying that they did not feel like that really—

HOWARD ROLSTON: Yeah, they didn't *feel* like it—

(Laughter. Cross talk.)

KATHLEEN ENRIGHT: Oh, you're saying we did not have the *proof* that you are—

HOWARD ROLSTON: Well, I don't think that they produced evidence for it. And you're right—they're big enough to be able to do something better, and hopefully they will from now.

KATHLEEN ENRIGHT: Yes.

ALBERT RUESGA: The Social Security Administration, I hear, was independently evaluated and got high marks. (Laughter.)

KATHLEEN ENRIGHT: Okay!

ALBERT RUESGA: They would be eating dog food without their Social Security payments, so who knows?

GARY WALKER: I would not answer your question generally, David (Roodman). But the one thing I do feel pretty confident of is, if we are talking about having more resources put into making organizations more effective, it is scary to me to think of that coming through the public sector. I just don't believe that there is the competence throughout the government. You hear a few people talking at the top who may know something about it, but that is not what a government is stocked with or the skills it has. I really do think that it is an issue that philanthropy is going to have to take a lead on or it will not be done well.

KATHLEEN ENRIGHT: Oh, but they first need to do no harm. That is the difference. I don't think that they should be doing any sort of capacity building, any of the government grantmakers. But they are some of the key contributors to the lack of infrastructure and underinvestment in people and systems that can produce the quality work that we're all hoping for.

JOHN FOSTER-BEY: I'm one of those federal people who may not be as competent as they should be.

(Laughter. Cross talk.)

GARY WALKER: That was a good opening, John!

JOHN FOSTER-BEY: Yes. Well, having said that, I want to say that this is—I mean, it's clear—a very interesting discussion. I find myself wondering, though, whether a lot of the controversy

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

around evaluation has a lot to do with our own desire to be able to do social engineering. I actually thought that one of the things you commented on, Gary (Walker), about not letting things fail—one of the reasons I think we are doing all of this evaluation is because we're not comfortable, we can't create a marketplace that allows things that don't work to go out of business.

I am actually increasingly a fan of not trying to make big bets on a couple of things, because I've also been a foundation grantmaker in some of the pretty big places and had responsibility for running some pretty good programs. I don't think we know as much as we think we know about what works. Somebody mentioned community development. Well, I am pretty familiar with that field and I will tell you, we really don't know what works in community development. We know how to build a unit of housing, but we don't know what really works in changing the community around. And so, I guess I am increasingly coming to—it's not a conclusion, because that would be much too strong, but I am coming at least to an openness to the idea that if we know a lot less than we think we know, maybe we should be a lot more humble about saying we can make big bets on things, and that we really should be asking, how do we create an environment that allows good things to at least have a shot at succeeding, and bad things to go out of business?

GARY WALKER: You are not one of the incompetent federal employees we were talking about.

(Laughter. Cross talk.)

GARY WALKER: For me, it's not a personal issue. It's a structural issue—what government is set up to do.

I think your comments are perfectly fair, and we certainly don't know in certain areas what to do. The only thing that troubles me is, I don't know how you get an atmosphere to allow more things to come unless we can start showing that a few things actually produce reasonable results—because I don't think that this is a country that is fundamentally sympathetic to this, although if you talk to people on the streets they are sympathetic to the brand names they know, Big Brothers Big Sisters and others. But a lot of the rest of this stuff means absolutely zilch to people.

KATHLEEN ENRIGHT: John (Foster-Bey), you brought up a great point about who is in the situation to do the social engineering. And there is a lot of hubris, and foundations often think, well, it's *us*, so I am going to hire a really smart guy who knows a whole lot about *whatever*—the environment, youth development—and I'm going to pay *him* to do the social engineering. Well, frankly, I am completely convinced that the people who are closest to the issues we are working on are in much better position to do the social engineering for the problem that they are trying to solve. It's a huge intellectual and programmatic shift for philanthropy and definitely for most governments to say, *we don't have the answers*; we have some supports, and we have some convening power, and we can help this group figure out the way forward led by the folks who are closest to the ground.

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

TERESA, Bill & Melinda Gates Foundation: I actually have a question coming from my background in development studies. I read a book written in 1970s by Arturo Israel about development projects observed, and it was about that same kind of debate. His big point was that you need to look at the type of project that is being funded. He talked about the concept of specificity; if the project has easily specifiable outcomes, then it is much easier to subject to metrics and evaluation. If it doesn't, you kind of have to take more of a leap of faith—if it's education or counseling—versus something like putting a laptop in every classroom or giving people vitamins. So, where do you think that falls into this discussion? Because we really haven't touched on that point yet.

GARY WALKER: This may not be responsive, but there is a whole class of programs out there called community change programs, community development programs—John mentioned this briefly. People know they want big improvements in the end in of a lot of things in the community, but it is so broad and specific that, really, evaluation has not played much of an important role in that field because you can't really get a handle on what it is they're doing. So there are fields like that.

KATHLEEN ENRIGHT: How I think that this plays into the conversation is, when you're thinking about what it is you want to learn and how you might go about evaluating a piece of work, it kind of goes back to the simple-complicated-complex, right? A logic model and a theory of change are pretty straightforward, good things if what you're trying to do is raise participation in the arts by people in a certain subcategory— stay-at-home moms, maybe. But it's not at all the right kind of approach if you're trying to counter racism. These are too high-concept, more tactical. So, one of the things that has happened is, a logic model and a theory of change have become a hammer for which everything is a nail. And it's not helpful. So developmental evaluations, evaluations that are participatory, where the people who are doing the work are highly engaged all along, where there's no hope of randomization or keeping fidelity to the model during the process, you *want* them to shift when they're learning along the way. It's where your head starts when you're deciding what the evaluation questions are, and whom you might engage to help you with it, and what your longer-term approach might be.

ALBERT RUESGA: I think your point it a very good one. I thought I heard Howard (Rolston) address it in part. It goes to the fact that many of us, myself included, who work in the nonprofit sector, are aspirational, and we don't often understand the practical limits to what our aspirations are. The connections between the work that we do and what we aspire to accomplish are sometimes very speculative, and they're not always very easy to see when you're talking about very broad areas of work, but much easier to see when you're much more narrow in your aspirations or make a lot more sense for the kinds of resources that you'll be able to put into a particular issue.

MARSHALL McNOTT: My name is Marshall McNott. I'm the recently retired president of the Los Angeles Mission for the Homeless. I want to say a couple of things that are not as esoteric that I think have a significant part in this whole thing that we've been talking about. I was reminded of it with Albert (Ruesga)'s comment about the church basement youth group. If the kids have a good time, that's enough—and that's fine *unless*, and this is a critical part, claims

P R O C E E D I N G S  
Metrics Mania? A Bradley Center Panel Discussion  
March 20, 2008

have been made to donors that the kids are also going to stay in school and stay off drugs. And that's the problem we get into. We make claims that we can't verify. *So just stop making the claims.*

KATHLEEN ENRIGHT: Seriously!

ALBERT RUESGA: Yes!

MARSHALL McNOTT: Let's keep it simple. There's a national homeless goal that is so absurd, and I've talked about it to mayors of Los Angeles, and they just shake their heads and say, well, we've got to do it that way so that the politicians will listen. The goal is to end homelessness in ten years. And I say, you know, you're going to end homelessness when you end poverty, you end mental illness, you end drug addiction, you have enough affordable housing for everyone who needs it—a total impossibility. But they stay with that goal, and some organizations continue to raise lots of money on it.

And the second critical piece—it's not that esoteric and another speaker mentioned it, but I'd like to take it a step further—is called humility. My experience, working several decades in the nonprofit sector, was that there's all too much side effect among nonprofits; we each think that we have the best program, so we don't need to talk to the others who are doing similar work *and learn from them*. We don't have the humility to talk to our fellow staff members, to the participants in our program. I've heard people say too many times, well, if he knew so much, why would he be homeless? Well, he might be homeless, but he might have some good ideas for one of our programs. So, the humility to listen, less arrogance.

HOWARD ROLSTON: I think that is a really very important point, that there is that kind of over-claiming that goes on. But I think that the other side of it is that there are some kinds of activities which only are of value if they lead to something else. One can think of other kinds of values, but if we set up a program to teach literacy, for example, maybe it's nice for people to be gathering together at that program or something like that, so it has some other benefit, but by and large for the public to invest in a literacy program that doesn't improve people's literacy outcomes doesn't make a lot of sense. It's not an investment. So, I think a good experience for a kids playing after school ought to be sold as that, but when other programs are, if you like, intrinsically extrinsic, and if they don't produce it, then it's not realizing an investment.

WILLIAM SCHAMBRA: One last question.

LISA SALES: I'm Lisa Sales, I'm the executive director of the Non-Profit Reporter. There was a lot of talk, and there is a lot of talk and a lot of panel discussions, but we're doing a lot of talking and not enough doing. And I'm wondering, can we package some things more broadly to establish a central repository, maybe to ask some tailor-made questions specific to the sectors and learn from what we gather? Maybe we don't, at the get-go, ask the right questions, but as we evolve over time we have a central repository that we can all learn from, and we can gauge, and we can change those questions and collect the data in a national central repository where we can learn what works and what doesn't so we can avoid the dares and we can avoid the scared-

PROCEEDINGS  
Metrics Mania? A Bradley Center Panel Discussion  
*March 20, 2008*

straights, the ineffective programs, or yet the harmless programs.

KATHLEEN ENRIGHT: Absolutely. And I think that several foundations are already taking this on. The Wallace Foundation in New York tries to ask and answer questions that their grantees individually couldn't answer, but will provide a breakthrough moment for them. And so they are trying to be a central repository on the issues that they care about: education, the arts. And they're sharing. The ultimate audience for anything that they do, from a research perspective, is the field that they're working in, not their board, not themselves or anything else. And I think that that's a reframing of the role of philanthropy in evaluation and in research.

HOWARD ROLSTON: Can I just say that in government, one of the issues is that all kind of government agencies have these what-works kinds of things, and you can look at them and they might contradict each other, but it all comes down to what the standard of evidence is for what goes in there. And if you look at like the Department of Education's what-works clearinghouse, they are very careful about what they do. They spend a lot of money on it, really trying to say to the public, to teachers, to whomever, there's really evidence that this works, or there's evidence this doesn't work. And I think that's a great service that they do. But then there are other places, where if somebody wrote in and said, gee, this is a good program, it gets on their thing.

ALBERT RUESGA: The repository isn't a panacea. It may be that you then implement very badly what it is that the documents in the repository suggest that you do.

WILLIAM SCHAMBRA: Let's thank our panelists for a terrific conversation.

(Applause.)