High Impact *Evaluations*

Exploring the Potential of Real-Time and Prospective Evaluations

Summary of a Workshop

IEG WORLD BANK | IFC | MIGA
Independent Evaluation Group

High Impact *Evaluations*

Exploring the Potential of Real-Time and Prospective Evaluations

A Workshop Conducted by the Independent Evaluation Group

Washington, DC
January 27, 2010

CONTENTS

Independent Evaluation Group, The World Bank Group
1818 H St., NW
Washington, DC 20433

IEG: IMPROVING DEVELOPMENT RESULTS THROUGH EXCELLENCE IN EVALUATION

The Independent Evaluation Group is an independent unit within the World Bank Group; it reports directly to the Bank's Board of Executive Directors. IEG assesses what works, and what does not; how a borrower plans to run and maintain a project; and the lasting contribution of the Bank to a country's overall development.

The goals of evaluation are to learn from experience, to provide an objective basis for assessing the results of the Bank's work, and to provide accountability in the achievement of its objectives. It also improves Bank work by identifying and disseminating the lessons learned from experience and by framing recommendations drawn from evaluation findings.

The findings, interpretations, and conclusions expressed here are those of the author(s) and do not necessarily reflect the views of the Board of Executive Directors of the World Bank or the governments they represent, or IEG management.

The World Bank cannot guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply on the part of the World Bank any judgment of the legal status of any territory or the endorsement or acceptance of such boundaries.

ISBN 13: 978-1-60244-192-7
ISBN 10: 1-60244-192-8

Contact: IEG Communication, Learning and Strategies (IEGCS)
e-mail: ieg@worldbank.org
Telephone: 202-458-4497
Facsimile: 202-522-3125

ACRONYMS AND ABBREVIATIONS

CEO	Chief executive officer
IDRC	International Development Research Centre
IEG	Independent Evaluation Group
IFC	International Finance Corporation
IFI	International Financial Institution
INTEVAL	International Research Group on Evaluation
GAO	Government Accoutability Office (United States)
GDP	Gross domestic product
MPL	Maximum probable loss
NAO	National Audit Office (United Kingdom)
NBC	National Broadcasting Company
OMB	Office of Management and Budget (United States)
RBS	Royal Bank of Scotland

All dollar amounts are U.S. dollars unless otherwise indicated.

Foreword

International development is undergoing a transformation driven by fundamental shifts in the global economic landscape. The 2008 financial crisis sent a shock-wave across the global markets and threatened to erase years of progress in development and poverty reduction in developing countries. It underscored the changing nature of the global architecture, a major aspect of which is the speed at which change occurs and the growing need for rapid and informed responses to potential and ongoing crises.

As evaluators, this means our approach to our work must undergo a sea change. As policy makers act on issues with very high stakes such as the global financial crisis and climate change, where the long term impact of ongoing actions can benefit from early feedback, we must be ready to provide an assessment of the likely effectiveness of their responses – even as those responses are being formulated. To seize this opportunity, evaluators need to revisit existing evaluation frameworks, respond to the uncertainties of the time and be willing to provide inputs that inform current and future directions. We need to work in real time so that our contribution is relevant, useful, and impactful. We need to generate findings that facilitate continuous learning and feed into a forward looking perspective.

In January 2010, the Independent Evaluation Group (IEG) held a workshop comprised of academics and practitioners of real-time and prospective evaluation techniques to exchange ideas and experiences. IEG's subsequent works on the World Bank Group's response to the global economic crisis are informed by the discussion at the workshop. We hope that this report – which includes a complete transcript of the workshop – and IEG's evaluations will help with your work as an evaluator or as a consumer of evaluations.

Vinod Thomas

Welcoming Remarks

Daniela Gressani, Deputy Director-General, Evaluation, World Bank Group

We are here today to discuss a topic that is relatively new and on which we all, I think, have a lot to learn, but a topic that is really becoming important. As evaluators, we have long recognized the need to make sure that our work has an impact, the greatest possible impact, and providing analysis in a timely way is a fundamental prerequisite for us to have a greater impact. I think that this is especially important now when institutions such as the World Bank Group have grown in size and in scope, and at a time when all the international financial institutions (IFIs) are struggling to respond to a financial crisis that has become an economic crisis, which, in turn, will require that we distill lessons and evaluate in real time.

This is what this day is about: trying to learn from one another how to improve the quality and the timeliness of our evaluation at a time when time is, in fact, of the essence. The value added of today's meeting is precisely to bring to the benefit of our own evaluation work on the World Bank Group the experience of important partners in other institutions.

SESSION 1: CONCEPTUAL ISSUES

UTILIZATION-FOCUSED EVALUATION: REAL-TIME AND PROSPECTIVE ASPECTS[1]

Michael Quinn Patton, Organizational Development and Evaluation Consultant

I am delighted to have the opportunity to be with you for this important discussion today, and am honored to kick it off. Let me remind everyone of some of the larger context. Tonight, President Obama will deliver the State of the Union Address, and had I known that when I prepared this, I would have called my presentation the State of Evaluation Address, but in that spirit, let me invite half the room to interrupt my presentation every three minutes with a standing ovation and the other half to boo and make rude remarks as I proceed to get us warmed up for this evening's adventure.

[The book] *Utilization-Focused Evaluation*[2] covers a great deal of our history, and I want to use that to talk about the state of evaluation as context for this consideration. The first edition of that book came out in 1978 and was basically reporting our findings on a study of use in the federal government, and the importance of the personal factor in how evaluations get used. The second edition, in 1986, brought together, from a lot of the work being done on use, the importance of intended use by intended users, being very clear about the purpose of any given evaluation and who it is for. In the 1997 edition, I introduced, as a field that I was coming to be aware of, the idea of process use, which is the way in which how an evaluation is conducted has an impact quite apart from the findings—things like capacity building, what gets measured gets done, the creation of logical frameworks and logic models for evaluation that begin to have an impact before any data are collected. And that has become a major theme of the last decade, which I think is quite relevant to real-time and prospective evaluation. The major new direction of the latest edition, which came out just over a year ago, was the challenge of evaluating under conditions of complexity. In a sense, in a thumbnail, that is some of the learning about the way in which the profession has emerged. That means that this session, and the direction that the Independent Evaluation Group (IEG) is going, are very much on the cutting edge of the larger issues that the profession faces.

PREMISES

Utilization-focused evaluation is a decision-making framework for enhancing the utility and actual use of evaluations. It begins with the premise that evaluations should be

[1] See Annex for full paper.

[2] Michael Quinn Patton, *Utilization-Focused Evaluation*, 4th Ed. Thousand Oaks, CA: Sage, 2008.

judged by their utility and actual use. Therefore, evaluators should facilitate the evaluation process and design an evaluation with careful consideration; everything that will be done from beginning to end will affect use. So a part of what I want to call to our attention is that realtime and prospective forms of evaluation have utilization implications: not just the timing of evaluation, but issues of credibility and quality and speed and all those things that are challenging the profession.

Some of what we have learned about use may be germane here. We have learned that use is a process, not an event, and that it needs to be facilitated. It involves an interaction, not just a report, to interpret findings and apply them. It involves training for use, not just the delivery of results. The intended users have to have some help in knowing what to do with findings. It is not apparent or natural to go from data to action and decision making, and use will mean different things for different evaluation purposes.

Evaluation is now part of an initial program design, including conceptualizing theories of change. Whether evaluators are present or not the very notion of theories of change has become so prominent that evaluative thinking becomes built into the program design process, and complexity is itself a theory of change about how the world works. The evaluator's role is to help users clarify their purposes, hope for results, and change the model. Evaluators can and should offer conceptual and methodological options. Evaluators can help by questioning assumptions. We play a key role in facilitating evaluative thinking throughout implementation as well as evaluation, and designs can be emergent and flexible, which is one of the challenges we are going to be talking about today, one of the new directions in evaluation.

For me, the big context here, my own bias about this, is that we live in a world that is increasingly driven by and paying attention to various forms of evidence-based practice. I like to say that evaluation grew up in the projects, testing models under a theory of change that pilot testing would lead to proven models, it could be disseminated and taken to scale. The search for best practices-- evidence-based practices-- remains one of the dominant, if not the most dominant, approach in much of philanthropy, in much of government and international agency funding. But what that comes up against is a fundamental debate, both intellectual and practical, about how the world has changed.

Whether it is through the top-down dissemination of "proven models" or a bottom-up adaptive management, this is a fundamental issue that, at the macro level of theories of change, is what brings us to issues of complexity. These are competing views about how the world is changed. Evaluation is a part of that debate because what we produce is going to be what informs both of these approaches, either the top-down dissemination of proven models or to be able to inform adaptive management, which is indeed real-time and prospective.

This also relates to an important distinction between dissemination of models and dissemination of principles. Best practice models yield recipes for exactly what to do, and the form of evaluation associated with that is fidelity evaluation. Is the model being replicated exactly as evaluated? Principles come out of bottom-up adaptive management, and when we generate principles and lessons learned those are not recipes. They have to be interpreted and adapted and applied within complex adaptive systems and contexts. That is a very different process than the high-fidelity replication of a proven model. Which means that the conditions that challenge traditional model testing evaluation, which I want to suggest has been and remains the dominant paradigm in the field and the dominant paradigm as I interpret IEG's work, the conditions that now challenge and lead us into this new direction are high innovation, rapid change, high uncertainty, dynamical, not just dynamic, systems.

EVALUATION, COMPLEXITY, AND DYNAMICAL CHANGE

Dynamical is a word in the complexity language that means ups and downs, not simply increases. Dynamic systems are on a pattern of increase or decrease; dynamical systems fluctuate in unpredictable and uncontrollable ways, emergent of factors in situations and overall systems change, all of which require and respond to adaptive management rather than a top-down, evidence-based, fidelity-driven approach to either implementation or evaluation.

Reminders, which we hardly need, but are part of conceptualizing our discussion of sudden change in massive uncertainty: 9/11, the Rwanda genocide, the SARS epidemic. When SARS hit Toronto, I happened to be working in Canada at the University of Toronto. There were ultimately about 40 people who died of SARS, and the economy of Toronto took a 25 percent hit from which it took two years to recover. The Wolfowitz scandal and resignation from the World Bank: I presume that was not an expected event. The global financial meltdown, which we are talking about today, the H1N1 virus, natural disasters like tsunamis and earthquakes. And closer to my own home, some of you will recall that on August 1, 2009, the bridge that was the main artery running through Minneapolis, suddenly collapsed at five o'clock in the afternoon, the main freeway that was the link not only for the Twin cities but for the entire state of Minnesota, and indeed the entire region. Ten days before this road collapsed, I was part of a group kayaking on the Mississippi River. We put our kayaks underneath this bridge and hiked up the bank to a coffee shop, came back down, we were cleaning up the river along that section. So when the bridge collapsed, I can assure you that it fell on clean ground. There was no trash to interfere with them later, but this has completely remade the transportation system in Minnesota, with huge reverberations that are still going on, an unexpected and uncertain event.

Evaluation's traditional comfort zone has been smart goals, controlled interventions anddefinitive findings -- traditional social science methods rendering major judgments. The emergent realities outside of our comfort zone that we are here to talk about are where uncertainty rules, where control is an illusion, and where complexity is the norm. Part of the issue is how to know what that territory is and what its implications are.

Many of you, I suspect most, are familiar with Nissam Taleb's important book, *The Black Swan*[3], in which he argues that the kind of events that I just went through, as highly uncertain and unpredictable with big implications, are actually much more common and much more dominant than people acknowledge. Indeed one of the extraordinary things about his 2007 book is that he predicts in great detail the global financial crisis, regularly described by economists and financial managers and gurus as an outlier event, and the reasons that it would occur. He argued that the the major reason the crisis would occur was because the entire economics and financial world was treating its likelihood as an outlier, outside of their probability estimates. He argues that black swans are common, they are definitive, and they are what control the world, not our normal activity. What goes on between black swan events is actually a temporary adjustment to the last black swan event.

EVALUATION AND STRATEGY

Let me introduce into this discussion Henry Mintzberg's work on strategy. Mintzberg is one of the major writers on strategic management. He is at McGill University. The *Wall Street Journal* has identified him as one of the 10 most influential management consultants of the last 30 years. He came out with a book in 2007 called *Tracking Strategies*[4], which is actually an evaluation book, although Henry did not recognize it as such until I met with him and told him that was what it was. But that book has 13 case studies of major multinational private sector organizations, and government and NGO organizations, that he has tracked over 20 to 30 years of what has happened with their strategies. And the picture that emerges from Mintzberg's work is that any organization begins with an intended strategy in a proposal, in a strategic plan about what they want to accomplish, and then they go into implementation, and the implementation of that he calls deliberate strategy, but every organization ends up having a part of that strategy that is unrealized, and then as they implement, there are new emergent strategies that end up as realized strategy.

So what he is saying is that high-performing organizations, in a five-year period, will begin expecting to go somewhere, and a part of that they will realize, but they will inevitably leave some things behind, and some new things will emerge, and where they end up in five years will not be where they thought they were going to be five years ahead. That is normality. That is also complexity. Now the implication of this is huge for evaluation, because our classic accountability model is to evaluate programs and projects on whether or not they ended up where they thought they were going to be five years earlier, and Mintzberg's work says no effective organization does that.

[3] Nissam Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*, 2nd Ed. New York: Random House, 2010.

[4] Henry Mintzberg, *Tracking Strategies: Toward a General Theory*. Oxford: Oxford University Press, 2007.

The challenges, then, are situation recognition and appropriate evaluation designs, and I am going to take you very quickly through a definition of complexity, and then Tom and others will tell you what to do about it. So I am just going to try to help define the territory of what it is that we are talking about. The context for this is research that is going on about expertise, the nature of expertise, artificial intelligence work around trying to model expertise, is that expertise does not consist of answers to things. Expertise is actually defined as situation recognition. What great experts bring is a knack for being able to understand what situation there is and the answers and responses flow from situation recognition. So we are talking about a contingency-based form of evaluation that is based on situation recognition, context sensitivity, clarity about who this is for, clarity about what it is for, matching methods to the situation, while maintaining criteria of credibility, meaningfulness, and timeliness.

To look at this through a complexity lens means that we are dealing with non-linearity, we are dealing with emergence, we are dealing with dynamical interactions, we are dealing with uncertainty, we are dealing with adaptation. What is this complex territory? Let me distinguish between simple, complicated and complex, and I have got a full paper in your packet that goes into this in more detail [see Annex]. It is also in a chapter in the *Utilization-Focused Evaluation* book, and it is a basis of a new book I have coming out in June that is entirely devoted to complexity evaluations. It is built around these distinctions, which I am going to run through very quickly.

We use a two-dimension matrix that my colleague Brenda Zimmerman developed based on work of Ralph Stacey out of organizational development. On the lower dimension is a continuum of how much we know about things, how to produce a desired result, a degree of certainty dimension. The vertical dimension is how much there is agreement on what to do and whether to do it. What we have here is a combination of these two dimensions that gives us a matrix of the interactions between degree of certainty and degree of agreement that defines different kinds of situations. Where there is a higher degree of certainty that we can produce an outcome and a higher degree of agreement that is called simple space. This is a descriptive term, not a pejorative term. It is not simplistic, it is simple. It means we know what to do, this is the realm of best practices, this is the appropriate realm of randomized control trials, this is the only place where that actually works, where you can do best practices. It is the realm of vaccines, it is the realm of polio eradication. The world has decided they want to eradicate polio, there is agreement about that, we actually know how to do it, and we are on the verge of doing it because it is in simple space.

Technically complicated things are things that have lots and lots of parts that you have to fit together that require lots of features. Launching the space shuttles is technically complex. Socially complicated things have lots of people involved, and the congressional analysis of the cause of the Space Shuttle disasters was partly technical, the O-ring and the foam, but was largely social, the culture of NASA, the interaction between the political people and the technical people. Socially complicated things are human rights agreements, environmental initiatives and the global financial situation, and socially complicated situations pose a challenge of coordinating many players.

So we finally get to the zone of complexity. Complexity is characterized by high degrees of uncertainty, we do not actually know what to do, and high degrees of disagreement about what the situation is, what ought to be done, and the politics of the situation. The farthest outside is chaos, which is best to avoid, but sometimes inflicts itself upon us, and the description of the eight days after Lehman Brothers failed, if you read it in *The New Yorker Magazine*[5], the 24/7 bringing together of the world's financial leaders and the world's bankers, is the best description of utter chaos that I have ever read. Absolutely nobody had any idea what was going on and were scared to death. This framework is being used by David Snowden, the former Director of Knowledge Management at IBM, who now directs a major consulting business called Cognitive Edge, and wrote a very widely disseminated article in the *Harvest Business Review* in November of 2007[6] about applying complexity to management, and given we are at the [International Finance Corporation] IFC and the World Bank, it is helpful to have a business kind of framing for this, which is why I am drawing upon Henry Mintzburg and people like David Snowden. Snowden's conclusion is that wise executives tailor their approach to fit the complexity of the circumstances they face, and what he is doing these days is training companies in how to deal with complexity mainly through real-time kinds of evaluations. That is his approach.

CONTINGENCY-BASED DEVELOPMENTAL EVALUATION

This brings us into a contingency-based developmental evaluation, applying these kinds of complexity concepts, matching the evaluation process and design to the nature of the situation to achieve intended use by intended users. A contingency-based approach beyond summative and formative, beyond static accountability models, to real-time, prospective, emergent action evaluation, adaptive evaluation, what I am calling developmental evaluation, as opposed to development evaluation, in the paper that is a part of my presentation. You will see that I distinguish both that all real-time evaluation is not complexity adaptive, and not all developmental evaluation is development evaluation. I make those distinctions.

I have identified five issues that are not unlike the issues that Tom Ling is going to take you through. Where I would leave you, based upon identifying and defining the realities of the world of complexity that we are going to be talking about, is the mantra for our time and for today that wise evaluators tailor their approach to fit the complexity of the circumstances they face. Thank you.

[5] James B. Stewart, A Reporter At Large, "Eight Days," *The New Yorker*, September 21, 2009, p. 59.

[6] David J. Snowden and Mary E. Boone, "Leader's Framework for Decision Making," *Harvard Business Review*, November 1, 2007.

EVALUATING IN UNCERTAIN ENVIRONMENTS: PROSPECTIVE EVALUATION AND SCENARIO BUILDING

Tom Ling, Head of Evaluation and Audit, RAND Europe

This paper builds very closely on Michael Patton's. But it does come from rather a different set of concerns and anxieties. The first is the experience of conducting real-time evaluation for the Department of Health, the European Commission and others over the last five or six years, and realizing that some of the most important things that come out of it are connected to the learning that took place and the changes that took place during the life of the project, and how important it is for an evaluation to track and learn the lessons from the changes that took place. Evaluating whether or not the original objectives were achieved can sometimes be less revealing than evaluating how and why the delivery was adapted to meet changing circumstances.

The second is a longstanding interest in scenario thinking and in thinking about whether, when you are faced with the kinds of uncertainties that Michael has been talking about, you can construct potential scenarios, plausible images of the future, in which you can test your strategies. At the back of my mind is the thought that there must be a way of linking that approach to real-time evaluation or ex-ante evaluation. A further thought is "What should the role of the evaluator be in this process?" We are used to thinking that evaluations should be both summative and formative but when and how should the evaluation itself become a driver of change?

For example, the work we do for the European Commission on Impact Assessment will typically ask us to construct three different approaches and test their effectiveness in the future. This typically involves presenting the European Commission's preferred approach, a "do nothing approach", and then there is the radical or extreme approach. We are then required to say which of these three options is the best. Many have got anxieties about how this approach to impact assessment is constructed. However, my point is not to question the details of this approach but to ask whether we should adopt a radically different approach to such ex ante evaluations. The thing that always occurs to me is that I would like to take a completely different approach and take the preferred approach and see how robustly it holds up in different plausible futures, as opposed to taking different approaches and seeing how they thrive in exactly the same future. This paper plays to that issue as well.

The third thing that has influenced this paper is working with an organization called INTEVAL (the International Research Group on Evaluation), where for many years we've been arguing that we need to move evaluation as a discipline away from major studies, typically at the end of projects, towards streams of evaluative learning, where evaluation is wrapped into ongoing events in a way that can support effective learning and contribute to accountability. In this approach the evaluator is more immersed in the process of learning and improvement (but carrying the risk that they may lose their impartiality).

Then the final factor influencing this presentation is, of course, that we undoubtedly are in turbulent times. Not everything will be turbulent, but many of the things that we as evaluators are trying to engage with will be more turbulent than heretofore. The important point here is that many of our evaluation frameworks help us to examine the costs and benefits of incremental changes, and to compare one standardized intervention against another. In contrast, evaluating complex and dynamic interventions requires us to look past the overt features of intervention and context and try to understand the deeper, more systemic processes at work.

So this paper builds very much on Michael's substantial shoulders, but it does suggest one particular way forward: exploring the relationship between scenario thinking and real-time evaluation. It is not a solution to all the anxieties raised in this introduction but it provides a pragmatic way for addressing at least some of them.

THE CASE FOR INTRODUCING SCENARIO THINKING TO REAL-TIME EVALUATION

I am going to look at three dimensions of the problem: deep uncertainty; evolving preferences and perceptions of utility changing during the lifetime of the project, and scenario planning. Mintzberg's questioning of old-style strategic planning, discussed earlier this morning, speaks to a world where projects and programs are purposive and include forward thinking and preparation, but evolve and adapt as practitioners learn and the world changes. In passing, we should also note that this makes identifying a single counterfactual even more complicated in this situation.

So with deep uncertainty, evolving preferences, and the absorption of strategic planning into learning organizations, there is a great need for a new approach to evaluation. Traditional monitoring and evaluation frameworks struggle to deal with projects that adapt or radically change their planned activities in order to achieve their original objectives. In other words they may keep the same objectives, but they actually change how they're going to achieve those. Secondly, many programs and projects quite rightly respond to unexpected changes in their boundary partners, who they may influence but cannot control. (Boundary partners are those organizations and groups who are a necessary part of the chain of causality linking the project to intended outcomes but who are not controlled by the project). These organizations whose behavior is crucial to the successful delivery of the project, may react in ways that were not anticipated, and, if so, the project may justifiably feel the need to adapt to these behaviors. And, thirdly, they seek to maximize utility: for example, you may have a program to reduce infectious diseases but divert resources to meet new needs resulting from natural disasters or civil war. Should the program manager be punished for that by your evaluators or should you be rewarded for showing flexibility and initiative?

Furthermore, most interventions are, in practice, self-limiting, and delivering continued long-term benefits requires multifaceted and evolving strategies (or sunset clauses and exit strategies). In turn, this requires non-linear, complex, and emerging evaluation strategies. Since most evaluations don't do this, most evaluation information is weak and fails to convincingly deal with attribution or accountability.

That's easy to say. It is a bit like saying to the caterpillar with arthritis, "I've got the solution for you, my friend, you need to become a worm." And the caterpillar says,"Great, how do I go about doing it?" And you say "Hey, I do the strategic thinking around here – your job is just implementation." It is quite difficult to actually absorb lessons we are learning as practical evaluators. The difficulties are numerous, but non-linear evaluations can become simply arbitrary and as shifting as a thing they evaluate. In other words, you may not really say a great deal, you just track a lot of changes taking place and finish up with a final report that might be summarized as "a lot of things happened, nothing much worked as intended (but some benefits were delivered), and there are no transferable lessons." We need to identify a set of agreed methodologies instead of reinventing real-time evaluation every time.

So we might think about real-time evaluation as a cycle of learning and accountability in the face of uncertain futures. Instead of the classic evaluation questions (what were your objectives, were they achieved?) we can ask a number of key questions about the capacity and skills demonstrated in dealing with complexity and change. We can ask periodically not only what has been done but also what is being learned? We can ask how the project equips itself to deal with uncertain futures. We might ask have you got robust ideas that hold up in different futures? Have you identified the different risks that exist in the areas? And there are risks. Have you got the skills that you might need to deal with those different areas? Are you monitoring the right elements in your environment? Are you identifying the key boundary partners that you need to influence in order to deliver on the program?

There is a related set of questions concerning how decision-making is devolved to those who have the best information and greatest capacity to exercise effective judgments. Is the program sufficiently adaptable? Have you got that capacity to adapt? Have you got recognition of and responsiveness to environmental signals in your world? How are your incentives working to avoid a program carrying on doing the same thing long after it had become sub-optimal?

In this context, evaluation becomes locked into a cycle of learning, supporting decision making, and demonstrating to others the reasons supporting the changes made. Evaluation material might then begin to look very different from the ex post evaluations we are used to seeing. They may take the form of annotated learning logs, for example. They may not even be done by professional evaluators.

WHAT IS TO BE DONE?

For all the growth in evaluation in the past twenty years it is not obvious that either organizations are better at learning or that we feel more able to hold organizations to account.

In the changing world we are describing today, what would happen if we thought less about evaluation reports and more about a stream of evaluation products? Or, even more heretically, we focused on evaluation activities rather than professional evaluators? The purpose would be to support well-founded judgments in the face of a changing world and applying lessons learned. It would also be to demonstrate to those holding them to account that this had been achieved. I am suggesting the production of evaluation products at key stages in the process of scenario-based learning. How has the project taken stock of their current situation, how have they identified the range of likely futures, how have they adjusted their understanding of the risks they face, are they still influencing the things in their environment? With this approach you can finish up with an evaluation-which rewards and explores and interrogates the capacity to act and respond at least as much as it addresses the extent to which you've achieved your initial objectives.

This would be one way—I do not at all want to argue it is the only way—of doing real-time evaluation in the face of uncertainty. It supports a creative response to the reality that there are multiple plausible futures that we face. It would, I believe, support good, helpful interim evaluations of progress that would be relevant both to the projects and to the wider community they serve. It can also be a way of including your boundary partners. It could help to build consensus about what those future challenges are. It can develop your ex-ante evaluations of capacities. It can provide an inclusive and supportive evaluation, and it looks at what have you achieved and how might you adapt, and evaluatees are not penalized for an inability to predict, but they are penalized for a failure to learn and adapt.

However, if you are going down that road, it does seem to me there are significant issues and problems. Some of those would resonate, I think, with IEG. When do you simply become implicated in strategic management? What do we really think about evaluations becoming agents of change?

It is the independence of evaluations, the dispassionate voice, which is in danger of getting lost. You would get very rich narratives, but you might lose accountability for performance against agreed standards. At the end of the day, public money, charitable money, or private money has been put into achieving public objectives, and people are entitled to ask were those objectives delivered on or not.

So just to recap, I think that the things that Michael has identified are significant changes for evaluators, and we are looking for different ways of reacting to that world. I think that the conceptual framework that Michael offers is very valuable and useful. I have identified one, I think, radical, way of working within that to try to build into our real-time evaluation something that takes at its heart the idea of the uncertain, the complex, and the need for adaptive, responsive but accountable organizations.

DISCUSSION

Daniela Gressani

One question that I have, and I think is really for both speakers, perhaps more for Tom, is the question of risk. When we are looking at things in the middle in real time, one of the things that we need to take into account is the possibility that things would not work as planned. I think that the testing of facts against alternative scenarios is part of this thinking, but how do we choose the right risks? How do we identify the downsides that we need to take into account when we, in fact, construct this scenario or more generally when we ask the questions that we need to ask? I mean, part of real time is that we do not know how things are going to work out before the full implementation of what we evaluate.

Hans-Martin Boehmer, Manager, Communications, Learning and Strategy, IEG

These were two very interesting presentations. Before I came to IEG, I was the head of Corporate Strategy for the Bank, and we actually invited Mintzberg for a seminar with our Board members. He basically gave exactly the slide that you just showed, and the response from the Board members was, excuse my language, but the Marion Barry incident was still fresh. This is how management wiggles out of accountability by saying, "Don't measure us against our articulated strategy, measure us against something else," but things change.

So the Board did not buy it, and in part the Board did not buy it because public accountability is seen as a big thing, and the Bank operates in a realm that is socially complex, where quite often what you actually do about the development problem isn't necessarily agreed on, and the Bank is seen as a rather contentious organization. I have a hard time figuring out what this means for independent evaluation. If you have some reflections on that, I would be very appreciative.

Gail Richardson, Lead Operations Officer, Europe and Central Asia Region, World Bank

You have kind of thrown my world into a different sphere, so I appreciate that. It is a very compelling presentation. I have two thoughts that came to mind, and one was this fundamental challenge that we already face in country capacity. We are telling countries the Bank does not actually do the evaluations, we give them the technical support and the resources to have that be done. So we have had this paradigm where we set the strategy, identify indicators, and now we are saying, yes, but we also have to be able to operate in this fluid environment, which is very real and very true. It throws the IEG evaluation of that original strategy into question in terms of the relevance of that if we are not supposed to be where we said we were going to be anyway.

The other thought that comes to mind, in addition to the challenge of country capacity, is that one of the drivers for change that I see as part of this complex environment is the demand by consumers and beneficiaries to provide feedback and the mechanisms to do that. We have got the ability to get data through cell phones now, much better channels of communication through email and faxes and etc., so it used to be, well, you go out mid-term and get that information, but now we are saying do not just accept that, do it, have it be a more dynamic process.

Nidhi Khattri, Senior Evalution Officer, IEG

My question is around this whole issue of mid-term or prospective evaluation. When projects do change, or public programs do change, do you have a set of questions you have actually used in assessing whether that strategy ought to have changed? In other words, not getting into any scenario planning or the actual content of the work itself, but some guiding questions as to whether strategy should have changed, on what basis it should have changed, and so forth. So it takes us a level higher than actually getting involved in the management of the issue.

The second question relates to whether in your own work you have come across programs or projects that have in fact changed rapidly, because public agencies take a long time to change and to deploy resources away from one set of options into something radically different, it takes a long time, and it is very difficult to do. So how do you judge that whole process? Thank you.

Stephen Pirozzi, Senior Evaluation Officer, IEG

I want to repeat my colleagues, thank you very much for your presentations. I have a quick question about project-level evaluation. If in the beginning we have a set of criteria or expectations or benchmarking for a project, and five years later we realize that everything has shifted or changed due to unforeseen events, does it then become an iterative process with the transaction team or management to reset those benchmarks? If so, does that compromise independence? How do those get reset so we can properly evaluate a project?

Michael Quinn Patton

Well, you have raised a lot of stuff and all of it very important. Part of this is about accountability and independence, so a couple of broad brush things. The notion of what gets measured gets done is the basis of a lot of performance results and performance management, the very kinds of things you are doing. That does make targets rigid, and it focuses accountability on where you ended up and where you wanted to end up when you started out.

Let us imagine that instead of making the target a fixed target, that in fact what programs are held accountable for is their adaptability and resilience, and that their responsibility is to document the basis of that. The way this wriggles out of accountability is that programs do not know how to document the changes that they are making, and document the evidence in a formal and systematic way about what they are seeing that leads them to adapt to what is going on.

We are acting like real-time evaluation is some new creation. That is how most businesses run. Businesses run on real-time evaluation. I am amazed that the World Bank and IFC, coming out of a research paradigm, ignore the way that businesses operate. They change constantly based upon customer feedback, based upon what's working and not working. They do not do five-year reports to find out whether or not their new program worked. They get real-time customer feedback and adapt. They evaluate whether there was a sound empirical basis for making the adjustments, and so a part of the way you maintain your independence, and it is an important independent function, is to look at the paper trail, and the logic, and the data that inform those decisions. Are people just shooting from the hip, or in fact are they reasonably tracking what is going on, getting feedback, and making adjustments on the basis of what is happening so there is a rationale for adjustments, it is not just willy-nilly? You, independently, can look at the basis for those adjustments and determine their reasonableness.

You cannot do it against a counterfactual. Tom and I may well disagree on this, but I think the whole notion of counterfactual becomes irrelevant under conditions of complexity. There are a million counterfactuals, so there cannot be a counterfactual. That is a mechanistic kind of thinking. What you end up doing, in classic Herbert Simon terms, is a satisficing judgment. Were the adaptations made reasonable given the nature of the changes? Was there a rationale? And did people themselves readjust their targets and do so on some reasonable basis? Can you track the path? A complexity-based evaluation is a map of decisions and alterations that show you, ala Mintzburg, where you ended up and why, and a judgment about the soundness of those decisions.

One final quick comment. This stuff is one of the top-down bottom-up tensions I was describing. I introduced into utilization-focused evaluation a new form of use that seemed to me to become dominant, driven by accountability concerns. It is what I call mechanistic use. Mechanistic use is the effort by policymakers to remove judgment from the system by creating artificial rules of action, like "three strikes and you're out." Like if you reach a certain test score in a school, the school goes on probation. No discussion of what that means, no discussion of context. Remove from judges making judgments, put it in the law: you do certain things, certain results happen.

Now the evidence is that prosecutors know how to game "three strikes and you're out" and are doing it. The schools know how to game No Child Left Behind. This mechanistic kind of accountability—of policymakers setting artificial targets and then holding people accountable for them—is a direction that is a reflection of mistrust in our political economy. It is very dangerous, it is very destructive, it is the opposite of adaptivity. Complexity requires judgment. It requires a fair judgment. It requires looking at the satisficing kinds of real world conditions that go on. And therefore independence helps look at that and make judgments about that, but it will not be a mechanistic, performance-based, number-based judgment. It is going to require independent auditors and evaluators to actually own their judgments, and the criteria for judgments, which are the reasonableness of adaptability as people respond to complexity.

Tom Ling

The counterfactual issue is that if you lose the counterfactual, you lose one arm of traditional evaluation, and you then need to think about how you compensate for that. I agree, in complexity, you have got an infinite number of counterfactuals, and so you need to think about how you manage that. To understand changes in strategy and as strategies evolve, one of the devices that I've used fairly successfully is a project diary, in which the project managers are required to maintain a six-monthly diary, which identifies key changes to strategy and why they made those, and lists the reasons why, and that has been a very effective tool I have found.

A small example would be an effort to improve the treatment of people who self-harm, particularly in an accident emergency. That project started off with one theory of change, and it was significantly transformed by the findings that emerged, but also by the fact that it involved users within the project itself, and it produced outcomes we had not anticipated that were really very interesting. So tracking those changes through the project diaries was one way in which we at least had some kind of written document that we could then point to and support our conclusion, which was that they would react and respond very effectively to new information as it became available.

I think also there is the question about risk and uncertainty and how we deal with the problem of risk, which is extremely important. I would make a distinction. Risk, which is a calculable thing, particularly in Anglo-Saxon approaches to risk; it is the chance of something happening multiplied by the impact that that would have, both of which are broadly quantifiable or scalable if not quantifiable. A lot of what we have been talking about is uncertainty, which is not quantifiable or scaleable in that way. The types of risks that you can address would be things like random behavior—try to model out what would happen with different forms of random behavior. Or if you have got inadequate information, which means that what you are doing is risky, you have got ways of managing that by collecting better information or analyzing the data you have got more effectively. But you still have got deeper uncertainties, which is really what we have been talking about, where conventional risk analysis will not actually help you to develop your strategy.

I would sharply demarcate where they are dealing with risks that they failed to identify but should have done from uncertainty that was either accommodated or responded to, which they could not predict, but where they should have known that there was a danger of becoming very like Rumsfeld. But there were risks, there were uncertainties which they should have acknowledged as part of their program. So, separate out uncertainty from risk.

And then there was the issue of the counterfactual. By and large what I have tried to do is to evolve contribution stories using John Mayne's approach. My first act is to say, why do you think what you are doing is going to make a difference, transforming that into the theory of change that is testable, and then trying to develop data around that. And the aim of the evaluation is not to get at certainty of effect, it is to reduce the uncertainty that the project manager and those holding them to account have. So it is a core of uncertainty where you can, by a series of evaluative activities, reduce the level of uncertainty about the effectiveness of the project or the program over time, but that core never reaches certainty. You are aiming to narrow down and reduce the areas of uncertainty and be quite explicit about what is still uncertain in the evaluation, which means judgment comes in. Thank you.

SESSION 2: REAL-TIME AND PROSPECTIVE EVALUATION IN PRACTICE

REAL-TIME AND PROSPECTIVE EVALUATION IN PRACTICE: THE EXPERIENCE OF THE U.S. GOVERNMENT ACCOUNTABILITY OFFICE

Stephanie Shipman, Assistant Director, U.S. Government Accountability Office

Creating a new program or policy – like any change – involves risks and opportunities. Forging a new approach, creating new rules and procedures, altering relationships between individuals and agencies, creates opportunities to fix problems with the old way of doing things but also uncertainty about future success. Evaluation-based program planning provides an opportunity to improve the chances of program success through incorporating (1) program features associated with success in the past, and (2) oversight mechanisms to provide timely corrective feedback on program performance. A systematic approach to these tasks helps the program manager minimize risk by ensuring a balanced, comprehensive analysis of the new program or policy that identifies unmet assumptions, builds upon existing evidence, and anticipates and counters threats to program success. The U.S. Government Accountability Office (GAO) is charged with providing objective information to assist congressional decision making. GAO conducts a wide array of studies of programs and policies, both prospective and retrospective. Today I will briefly describe our experience with two types of studies that directly aim to assist program and policy design—prospective evaluation and real-time evaluation.

THE PROSPECTIVE EVALUATION SYNTHESIS

Developed at the GAO in the 1980s, the prospective evaluation synthesis is a systematic method for assessing the likely success of a proposal by comparing a new program or policy's features and assumptions to existing evidence on similar approaches. It is intended for use when a new program or alternative approach is being considered; the most effective approach is not known; but similar approaches have been tried (and tested) in the past. The method begins with an analysis of the proposal which articulates:

1. the nature of the problem the program is designed to address;
2. a conceptual "logic" model of the mechanisms by which program activities are expected to "fix" the problem; and
3. an operational model of what resources are required or assumed to be available.

After assessing the proposal's assumptions and internal consistency, data are collected, reviewed and synthesized to assess the quality and extent of evidence supporting the proposal.

GAO's Assessment of Teenage Pregnancy Program Proposals

In the 1980s, births to unmarried teenagers were rising alongside concerns about the associated negative social and economic consequences for these teenagers and their children. In the absence of a federal program specifically targeted to this problem, several legislative proposals aimed to create new programs to prevent teenage pregnancy or its economic disadvantages for young parents. GAO was asked to provide information on:

1. the extent of the problem;
2. the effectiveness of programs for preventing teenage pregnancy and for providing related services to pregnant and parenting teenagers; and
3. the implications of this information for structuring new legislation.[7]

To provide structure to the analysis, GAO selected two maximally different legislative proposals, from among a dozen being considered. Then, for each proposal, we categorized the strategies they took, including the types of services, locations, and populations they targeted. We then described each proposal with conceptual models that articulated the mechanisms by which program activities were expected to result in desired outcomes, and operational models that depicted the specified organizational arrangements.

To assess the promise of these conceptual and operational models, GAO reviewed research on the size and scope of the issue (to estimate the population eligible for each program), summaries of research on the antecedents and consequences of the problem (to compare to the conceptual models), and evaluations of similar service projects conducted at the state or local level. Evaluation studies were first screened for research quality, and then their results were summarized by program strategy and type of service for each desired health, education, and income-related outcome. To assess the operational models, we also reviewed the evaluation literature and a previous survey of program administrators to identify challenges to and solutions for operating these types of projects.

Lessons Learned

As you might imagine, we discovered that the success of the prospective evaluation synthesis method is highly dependent on the availability of good quality studies of prac-

[7] GAO, "Teenage Pregnancy: 500,000 Births per Year but Few Tested Programs," GAO/PEMD-86-16BR, July 1986, p.7.

tices that are similar to the target proposals, and have been used with groups similar to the intended population. Although similar teenage pregnancy programs had been evaluated before, flaws in their research designs and lack of data on long-term benefits limited our ability to identify “what works” in reducing the negative consequences of unmarried teenage parenting. Thus, there was little direct “hard” evidence on which proposal’s conceptual model – the comprehensive services or simpler approach – was more likely to be successful in achieving the desired outcomes.

On the other hand, evidence on the difficulties in implementing prior programs suggested one should keep the program’s administrative procedures fairly simple. Here, a lack of evidence on effectiveness did, nevertheless, clearly lead to a policy recommendation. Since there was no evidence that the more complicated comprehensive service model was more effective than the simpler model, there was no support for requiring adoption of the more complex model that would most assuredly be more difficult and expensive to implement.

Finally, the lack of clear evidence of effectiveness suggested that Congress might want to hold off on creating a new national program. Instead, they could consider creating a small demonstration program to carefully evaluate alternative service models in order to learn what works for future dissemination. That is, a small, targeted program with built-in feedback on performance can minimize current risk while also reducing uncertainty of success in the future.

REAL-TIME EVALUATION

After frustrating efforts to evaluate the effectiveness of programs only to discover that they had not actually been carried out as designed, program evaluators now expect program implementation to be evaluated before – or as part of designing – an effectiveness evaluation. An implementation (or “process”) evaluation assesses the extent to which a program is operating as intended, that is, conforming to statutory and regulatory requirements, program design, professional standards or customer expectations. It may address issues such as the appropriate and efficient use of resources, the quality of products or services, or the extent to which the targeted population is reached. While they could be undertaken at any time, implementation evaluations are typically conducted early on to identify and respond to emerging problems in a timely fashion. Real-time evaluation in the foreign assistance field has been described as a typically rapid process evaluation of a relatively brief initiative (several months long), intended to provide feedback to guide corrective action.[8] When interventions are this brief, it is probably especially important to draw on prior evaluations in program design and obtain rapid feedback.

[8] Maurice Herson and John Mitchell. “Real-Time Evaluation: Where Does Its Value Lie?” *Humanitarian Exchange Magazine* 32 (December 2005) www.odihpn.org/report.asp?ID=2772

GAO's Real-Time Assessment of Recovery Act Implementation

In early 2009, the American Recovery and Reinvestment Act[9] authorized an estimated $787 billion in new federal spending and tax provisions to respond to what is believed to be the Nation's most serious economic crisis since the Great Depression. The Act has an array of purposes: to create jobs and promote economic recovery; assist those most impacted by the recession; invest in transportation and other infrastructure to provide long-term benefits; and to stabilize state and local government budgets. Experience with other large federal spending initiatives has found that risk for fraud and abuse grows when billions of dollars go out quickly, eligibility requirements are established or changed, or new programs are created. Thus, both Congress and the Administration desired to ensure transparency and accountability in use of those funds to avoid waste, fraud and abuse. As one piece of the built-in oversight framework, the Act mandated GAO to, among other things, conduct bimonthly reviews of states' and localities' use of Recovery Act funds and approaches taken to ensure accountability for those funds; to assess whether the funds are achieving the stated purposes of the Act; and comment on the estimates of the number of jobs created and retained by recipients of Recovery Act funds.

Since March 2009, GAO has been collecting longitudinal data on the actual and planned use of Recovery Act funds in 16 states and D.C., which were selected to represent two-thirds of the U.S. population and two-thirds of the intergovernmental grant funds.[10] GAO also collected data on grant making and monitoring activities from six federal agencies overseeing Recovery Act grant programs that have begun disbursing funds to states or have known or potential risks. GAO assessed the reliability of the estimates of jobs created and retained through review of federal guidance and federal and state quality review procedures, and analysis of recipient data submitted to Recovery.gov.

Lessons Learned

From the start, GAO's reports (April 2009) provided valuable nationwide information on the uses and tangible benefits of Recovery Act funds at the state and local levels.[11] For example, GAO clarified that much of this state and local spending would not occur until 2010, and a majority of the initial grants went to state Medicaid programs of health care for the poor, elderly, and persons with disabilities. States reported using these funds to maintain Medicaid eligibility and benefit levels and cover increased caseloads due to the recession, as well as to offset state general fund deficits, thereby avoiding layoffs. The bimonthly reports also provided insight into the interaction of federal and state rules and processes that could – at the least – delay achievement of program benefits. For

[9] PL 111-5, Feb. 17, 2009

[10] GAO, "Recovery Act: Status of States' and Localities' Use of Funds and Efforts to Ensure Accountability." GAO-10-231, Dec. 10, 2009.

[11] GAO, "Recovery Act: As Initial Implementation Unfolds in States and Localities, Continued Attention to Accountability Issues Is Essential."GAO-09-580, April 2009.

example, in many states, legislative authorization is needed before the state can receive and/or expend funds or make changes to program rules. In some programs, the twin pressures for accountability and speed created difficulties. For example, by November, one-third of local public housing authorities were not on-track to spend funds for capital improvements within the allotted 12 months. This was due, in part, to large grants that led to more, and more complex, projects that required additional design work and clearances; and, in part, to additional federal monitoring of a small number of local authorities with troubled procurement histories.

Early monitoring and reporting can – and did – identify important issues to correct while funding is still being disbursed. (As of late November, three-quarters of the approximately $280 billion for programs administered by states and localities, had yet to be paid out.) Some GAO recommendations have already been acted upon. To respond to states' lack of funds for their new oversight responsibilities, [the Office of Management and Budget]OMB provided guidance on how to obtain some cost reimbursement, while additional funds are sought from Congress. To help states coordinate the various Recovery Act funding streams, OMB now requires federal agencies to notify state recovery coordinators of any awards made in their jurisdiction. To improve the credibility of recipient reports of jobs created or retained, OMB and federal agencies have worked together to improve guidance and conduct outreach, and they have re-examined their quality assurance processes after the first round of recipient reporting.

In particular, GAO recommended modifying and leveraging an existing oversight mechanism – the Single Audit Act – in order to simplify and consolidate some of the separate federal agency oversight requirements. To reduce duplication and fragmentation in federal oversight of state and local execution of numerous federal programs, the Act encourages reliance on periodic consolidated audits of these agencies' fiscal and program management. However, to ensure timely and efficient feedback on Recovery Act operations, GAO recommended accelerating the audit reporting timelines, applying audit requirements to some small but high-risk programs, and considering lifting these requirements for some low-risk programs. OMB is currently operating such a pilot project in several states.

Finally, this level of scrutiny of an unprecedented large, multi-agency initiative requires vast resources. GAO obtained special authorization for temporary hiring that allowed us to field audit teams across the country, in addition to our ongoing work. GAO also worked closely with federal agency Inspectors General, state auditors, and the Recovery Accountability and Transparency Board to share information and audit findings.[12] Although GAO has reviewed internal controls in new programs before, the bimonthly reporting cycle has strained the audit agency's capacity. Bimonthly report-

[12] The Board, including many agency IGs, reviews the processing of contracts and grants, reports quarterly to the President and Congress, and is charged with reporting any potential problems requiring immediate attention.

ing is highly unusual and burdensome for an audit organization that devotes significant resources to validating data, findings, and conclusions. Nevertheless, this type of comprehensive analysis - which draws on lessons learned over time in the areas of fraud prevention, contract management, and grants accountability – will help control risk and increase the Recovery Act's chances of success.

Program evaluation – unlike research – is primarily conducted as an aid to decision making, and oversight agencies, in particular, aim to help policy makers manage risk and opportunity. Thus, as evaluators we seek to marshal credible evidence on how well programs have been performing and draw inferences about what we can reasonably expect in the future, based on available information. Evaluators need not be forecasters to be able to recommend ways to limit risk and encourage program success.

THE UNITED KINGDOM RESPONSE TO THE CRISIS: EVALUATING IN REAL TIME

Philip Airey, National Audit Office, United Kingdom

The National Audit Office (NAO) is, first and foremost, the United Kingdom government's financial auditors. We certify the accounts of many public bodies in the UK, but we also have a statutory responsibility to report on the value for money with which resources are used by the UK government. We produce around 60 reports a year. My presentation is about two of those reports and a few more to come over the next few years related to the financial crisis.

NAO STUDIES

We have published two reports so far. The first, on Northern Rock which was a relatively small mortgage bank based in the north of England, was produced early last year. I think one of the presentations earlier talked about complexity and chaos. This was a case of the British public verging on chaos. Many people thought there was some danger of losing their deposits in this bank when it got into trouble, so they formed an orderly queue outside each branch to withdraw their money.

We published a second report towards the end of last year. This report deals with the program of projects that has been put in place since Northern Rock and it is a "mapping" report. It is very much a non-evaluative report. It sets out what has happened and why, and positions the NAO for future evaluations over the next few years. For instance, we are conducting a program of work at the moment, looking at a large insurance scheme for one of our major banks. The scheme was being put in place when we did the last report and now that it is up and running we will examine it. We will also report on the unwinding of the measures as and when share stakes are sold and guarantees lifted. So this is very much a real-time evaluation, a set of real-time evaluations for us.

Why is the NAO interested in this? It is an extremely complex situation, with enormous risks for the UK taxpayer.

What are we doing? Well, two of our aims in doing this work are transparency and accountability. Many taxpayers in the UK are unsure of what's being done and why. Up until these reports, there had been little accountability to the UK Parliament for what was done. So those two reports by the NAO are the beginnings of a process of transparency and accountability.

First we had to define the scope of our work. If you are going to do an evaluation, you have got to think carefully about what you are going to evaluate. The first problem we had when putting together this piece of work was the question: what are we going to look at? We had to be careful here because there was a lot going on. Financial regulation in the past has not been part of the NAO's audit responsibilities and we do not examine the conduct of monetary policy. What we did cover, though, was the development and implementation by the UK Treasury and others of a whole series of support schemes. We are not looking at other, wider policy areas. You need to be quite careful as it's a huge area and very complex. I am not saying that we will not look at that in the future, but these reports are all about implementation by the Treasury of a series of schemes to deal with the financial crisis.

We also had to have a clear message very early on, and in all our work the first question we ask is: if the government spends money, did it need to spend all of that money in the first place? It's all very well having a fantastic project, but if you don't need to do it, there's no real need for it, well, why bother? We had to have a clear message, especially in this second report, that "do nothing" was not an option. If nothing had been done then chaos similar to that seen when Northern Rock got into difficulty would surely have ensued. So we had to get that question out of the way, but then comes a question that really does concern us: was all this value for money? £850 billion is a staggering sum of money. The NAO has never looked at a program of projects involving such a huge amount.

The program is made up of a whole series of schemes. Much of it is guarantees and insurance, both across the system and for individual banks, but there is some direct expenditure in the mix, some share purchases, especially in two of our largest banks, and some loans as well, to a whole series of smaller bodies and organizations. So in total there is just over £100 billion in direct net expenditure so far. We have got around £14 billion back, which is a start, and one of the things we will do is keep a scorecard as we go along, the cash out the door and cash in the door, in our future reports.

One of the first things we do when we launch an evaluation is to ask an organization: what is your aim and objectives for this project? So what were the objectives? Primarily to protect the financial system, protect depositors' money, those are the first two, they're not mutually exclusive, and the third, ensuring continued lending to creditworthy borrowers, is also about financial stability, perhaps a bit more of a long-term objective. Those top three are absolutely key and the final objective -- the one that interested us most -- if you have got to protect financial stability, what are you doing to protect the taxpayers' interest?

FINDINGS

So let us start off. Did they get the basics right? Past crises around the world were examined and there were generic solutions, but there was very little time to work out detailed plans. The UK authorities did not have adequate legal powers when faced with the crisis at Northern Rock, and their contingency planning for such an event was not up to date.

Resources: When we looked at the Treasury, about 17 staff were responsible for overseeing financial stability in 2007, so very few staff available -- very few staff with the skills that would be needed.What we've seen is a heavy reliance on the external advisors. There's a plentiful supply of investment banking advisors in London, as you'd expect, and they were brought in very quickly when needed.

Timing: Careful thought was given to the scope of what could and could not be done, from the do nothing option, right up to full nationalization of the banking system. The thinking was always around what would be a proportionate response? What's happened in the market this week? What would be a proportionate response? So we were satisfied that the taxpayers' interest had been protected in the sense that the schemes put in place were proportionate. The details of the schemes were worked out quickly after they had been publicly announced, which helped in producing a direct effect on the markets, and we could see that happening from all the published market numbers. But this was never a simple cause and effect relationship.

Performance measures: Another question we always ask when conducting evaluations is: you're doing this project, have you got a set of performance measures that will tell you when you've achieved your objectives? If there aren't any, we will try and develop a set of measures. At the start, such performance measures were underdeveloped. They are developing some now. However, in looking at this and evaluating it all, we have to bear in mind that this was a crisis situation. Nobody expected this to happen. Ultimately, the UK authorities did a pretty good job, and our reports say that.

Financial stability: Was it maintained? It was. No disorderly failures, no losses of deposits. Commitments are now in place to encourage lending to creditworthy borrowers. In evaluation terms, the success in meeting lending commitments is on our agenda. Financial stability was maintained, but when we tried to put a bit more of evaluation into this, when we looked at various market indicators of solvency, of liquidity, it was very difficult to isolate the effects of individual measures. For instance, the Bank of England has reduced interest rates to half a percent, and is now buying high quality assets. That also had an effect.

When we published this report we said that all the indicators that we could find were looking good, they were all heading in the right direction. We were not actually sure whether it was because of the individual schemes that we were looking at or whether other things were happening, and whether actions taken in other countries -- in the U.S. in particular -- to deal with the financial crisis were having a direct impact on what was happening in the UK. So that is all very complex, and we have not been able at this stage to cover that in our report. I make no apologies for that, it is just too complex for one of these reports. Perhaps in some future work we will come back to it.

Taxpayer protection: A big question for us. So far, so good, no guarantees have been called. They have had the effect intended and there is fee income coming in, based on market prices. That is one thing we did look at. Were they charging these banks for this support? They are and it's based on market rates.

Insurance: Again, the government undertook extensive due diligence before getting into this. There is a question about pricing there, which we are looking at. For various reasons the scheme could not be priced at market rates, but we will explain that in the next report.

Share purchases: These were done after extensive stress testing, so were proportionate in that they only did the minimum needed. We will come back to that when the shares are eventually sold. We will do further reports; again, we will have that scorecard in the background as well.

The lending to various organizations: Over-collateralized and priced at market rates, so taxpayers are protected and there is a lot of follow-up work for us over the next few years. Have we been able to fully evaluate? Actually, I do not think we have in practice been able in the fullest sense to evaluate this program and we will not be able to fully evaluate until about two or three years' time, when the measures have wound up. So far, so good. A bit of real-time evaluation for us, and it is a work in progress.

REAL-TIME EVALUATION IN THE INDEPENDENT EVALUATION GROUP: ASSESSING THE WORLD BANK GROUP'S RESPONSE TO THE GLOBAL CRISIS

Ismail Arslan, Senior Evaluation Officer; Daniel Crabtree, Evaluation Officer; Ali Khadr,Manager; Marvin Taylor-Dormond, Director; and Stoyan Tenev, Chief Evaluation Officer, IEG

The main catalyst for this work goes back to September 2008, and that was the collapse of Lehman Brothers. There was immediately a run on stocks globally. It spread very quickly to the developing world, not just financially, but also economically and socially. We have seen already quite an uptick in unemployment, and poverty levels are increasing. In short, it was quite an ugly scene that emerged, and we are still feeling a lot of the consequences of it.

THE WORLD BANK GROUP'S RESPONSE TO THE CRISIS

What happened here on Pennsylvania Avenue or H Street, where the Bank Headquarters are, I think we can categorize as an element of surprise when the crisis hit. There were some warning signs that the crisis was coming, but the Bank Group was initially focused elsewhere. The Bank was looking at the food crisis: food prices had increased rapidly in the previous 12 months, so it was handling that. And IFC was, in the first instance, concerned about ensuring that it had profits coming through in the current year so that it could make further investments. Its capital was rather constrained.

There was at that point a search for lessons and direction. Where do we go with responding to this crisis? This is where IEG comes in as part of a multifaceted story. We reviewed the lessons of previous crises very quickly. We looked back at 20-something crises from the 1980s and 1990s and reported to the Board on those lessons. In December, some new crisis initiatives were launched, covering a number of aspects, such as trade and infrastructure. There were announcements about new lending that would be carried out over the coming years, and some objectives finally crystallized in the spring of 2009. So there was a direction that IEG was helping to influence by looking at the lessons of the past.

These new realities—the doubling of lending, the fast tracking of lending, the possibility of greater impact, but also on the other hand the greater risks that come with the additional lending and speedy lending—are highly complex, highly uncertain prospects. That really ought to be the case for evaluation getting involved early on, on a real-time basis, so that we can promote learning from experience as the crisis response is being implemented.

Results are more important than ever, and the resources, of course, are constrained so there is less ability to carry out self-evaluation, and also an independent perspective is important. And if we wait, it is going to be too late to influence the direction, to provide learning, and to change course if things are not going the right way. Also, evaluation is important, of course, for accountability.

IEG'S EVALUATION WORK

Our approach covers the whole Bank Group, a joint effort, including IEG-IFC, IEG-World Bank and IEG-[Multilateral Investment Guarantee Agency]MIGA. It is a phased approach, looking first of all globally, at what is happening with the response, and then drilling down into country cases, sequencing the outputs that we will deliver, and updating on a live basis. For example, we finished a report in November, which we submitted to the Commitee on Development Effectiveness (CODE) of the Bank's Board, and then we had an informal briefing with the Board of the Bank in January. The report had the data until the end of the third quarter of 2009. The briefing contained data through to the end of 2009. So we used data as close to the current day as we could, and then because of the need for speed, used some less formal processes for delivering interim products, to brief the Board, to elicit management feedback for internal clearance, and for quality assurance.

There are a few challenges that we have seen already in carrying out this work. Firstly, arguing the case for doing real-time evaluation to the Board and management. The Board and management were accustomed to IEG doing ex-post evaluations. Conducting an evaluation in real time was in many ways precedent-setting. Second, data, given timing, is of course incomplete, so we do not have all the data on outcomes or impacts. Third, results frameworks are lacking in many cases, and there is a lack of baseline data and monitoring. Fourth, the challenge of balancing speed and quality in a real-time evaluation. Fifth relations with management, the point that Tom Ling was making before about playing a judicious independent role without getting in the kitchen, and, finally, timing our outputs so that they have the most utility, so that they will be well received.

How have we sought to address some of these challenges? Well, in making the case for real-time evaluation, we made the promise that feedback would be timely, that we would be able to offer learning. References to practice elsewhere, for example in the NAO and the GAO were very helpful in that they were carrying out that work at that time. Furthermore, the uniqueness and magnitude of this event argued for a case for new approaches. Regarding the incomplete data, we cannot deal with this fully given the timing, but we were able to pick off early aspects. So we could consider how relevant is the response? How well designed is the response? How is it going in the first year, factoring in lessons of past crises, looking at interim indicators, and focusing on what is coming out as well as what is going in, on actions and processes, and maintaining frequent contact with operations to get the latest data to be able to update on a live basis.

We carried out country visits as soon as we could, with a prioritization on the region that was initially hardest hit by the crisis.

To ensure quality, we held regular, high-frequency meetings of the team, but also had a steering committee, which cuts across the IEG to provide that guidance to the core team. We kept close engagement with management and, with departure from business as usual in terms of the informal exchanges, and understanding on the part of management that there is a need to do something a bit differently this time.

Timing for impact, we are looking ahead at what is coming, particularly the spring meetings at the World Bank Group. When does the Board want briefings to inform their thinking, their decision making, listening to what might be useful to management?

Where have we reached so far, what have we delivered? We did the notes on the lessons of past crises at the end of last year. Within the last 12 months, we have delivered our notes on the first year of the response, and briefed CODE just a couple of weeks ago. Feedback has been very positive that the work has been relevant and useful. We are victims of our own success at the moment: there is a demand for more, both deeper work and broader work, which poses some challenges, especially over the next couple of months, in being flexible in the use of our resources, paying special attention to interaction with management, and ensuring that we manage expectations. There are going to be some tradeoffs in that we have a relatively small amount of time to do the work, so we have to say no to some aspects.

FINDINGS AND INSIGHTS

At this point let us emphasize that we have passed these much more as descriptive insights than judgmental insights, because this is an ongoing study, and we have not as yet delivered any kind of formal report. By the spring of 2009, the Bank Group had clearly articulated its objectives regarding the crisis response, but there was not a clear sense of what would constitute success or otherwise. Now whether it is reasonable or not to expect such clarity is entirely another matter.

In terms of implementation the last fiscal year has been a year of historically high lending, even though it is a modest amount relative to the financing gaps. Of course, the issue of to what extent you catalyze the flow of funds comes up there. There has been a stronger poverty focus in the response operations than had been the case, for example, in the East Asia crisis, but the issue is how to sustain that. And IFC made a quick response in terms of structuring initiatives, but the problem really has been in implementing some of these initiatives.

In terms of initial lessons and results, prior country engagement matters for both speed and quality.This is a time of historically low conditionality in World Bank operations and much more country ownership. So what does that tell us about the importance of results frameworks to structure things and ensure sustainability? The private sector platform, terms of structuring initiatives, again, is great, but in terms of implementation it has been somewhat feeble in some areas, and that might have meant missed opportunities.

Issues going forward, and these are really things at this point that we identify as wanting to keep on the radar screen: Results frameworks, the importance of trying to use those in World Bank operations. The issue of responding and channeling the financial flows where they are most needed. Again, this is akin to the counterfactuals debate and has parallels with the issue of how long is a piece of string, but leveraging the unique World Bank Group reach, contingent capital arrangements, setting up delivery platforms and protocols ahead of time so you can structure a response in some established framework, or at least ensure that people know the rules. On quality of impact, you have got to worry about the fiscal sustainability of clients, strengthening the poverty focus, supporting growth reforms, delivering on the private sector response and, of course, not forgetting about long-term sustainability issues such as climate change and environment.

Now, in closing, let us to go back to something that Philip Airey said about the definition of crisis in the UK—having people lining up outside banks, and it really is so true that that happens. But you know, even in the stoic UK there is an almost daily event, because it rains almost every day in Britain, that throws off even the most stoic and determined British queuer, and that is the following: At a bus stop, when you are queuing and waiting for the number 28 bus, let us say, and you see a number 28 bus coming, and you think, “Great, here I am number 20 in line, and I’m probably going to get on the bus.” Then you look again and you see another 28 bus, and a third 28 bus, and sometimes even a fourth 28 bus, because of course all of them come at once, and then they don’t come for half an hour. So what happens with that is immediately the British system of queuing breaks down. You can tell the little old lady next to you is wondering, ”Should I make a run for the last bus, or should I try to get on the first one which is closer?” Of course, the last bus is going to be the emptiest, and so there is a big reward in terms of getting there, but it is further away. The parallels between that and the complexity of what evaluation has to look at are not lost on us. Thank you.

DISCUSSION

Roland Michelitsch, Chief Evaluation Officer, IFC

I have a question to all of the speakers. You mentioned that it is very complex even to have a model of how the interventions are really going to feed through the systems and how they interact with each other is very difficult. So how did you actually address that in terms of setting up in the evaluations where it is very difficult to attribute in complex situations? What is really attributable to a specific intervention? Then, secondly, we talked a little bit about the importance of what is your without project scenario, and obviously the more dire your without project scenario is, the better your with project scenario performs—the financial sector did not collapse or whatever you use as the without the case scenario.

For example, I know IFC best because I am from IFC, and when I look at past IEG results, and this is IEG data not my own data, I see that those projects that were actually approved during crisis situations tended to perform very poorly. If we went in right afterwards, we actually got very good results, and so how do you judge in that context the fact that IFC was focusing on the portfolio? Arguably, if you construct a without projects scenario, maybe all of these client companies would have gone under and our [Maximum Probabilty Losses] MPLs would have shot up and really constrained our ability to do something in the future. And on top of that if we had just pushed out money really fast, maybe we would have gotten, like we did in the past, really very poor results. So how do you factor that into actually the messaging that comes out of that? I really liked in the NAO case, but putting that in context while remembering this was a crisis situation and just being able to react very fast and so on means that sometimes you will have to make some trade-offs, speed versus quality.

Hans-Martin Boehmer

I would like to hear about the experiences as to how the real-time evaluations have been used by the management to maybe make some modifications or changes in their programs. I think GAO did mention that there were some views, but it will be useful to know how it gets used by the management.

Marvin Taylor-Dormond

I just wanted to hear a little bit more from Stephanie and Phil about the way you are dealing with results. Attribution is truly an important issue, but before attribution is measuring results. That is a fundamental, a key issue in our case, and it is precisely what has been behind one of the foundations of ex-post evaluation, because we argue that in development results take time and then only ex-post evaluation, five years after the project has been dispersed is the right way to do it. That would argue against real-time evaluation. My argument has been that is like saying that in a hospital the emergency department should not develop a results framework because that belongs to the long-term care section, and so obviously is incorrect. There is a different results framework in the emergency unit from that of the rest of the hospital. The problem here has been that the units have not been used to developing this type of results framework for emergency or crisis situations. So I just wanted to hear a little bit more on that. By the way, I just saw, Phil, that you said that you did not venture much on results, but were you courageous enough that there is a section of your report in which you are determining these results?

Tom Ling

Just very quickly, before coming here I feared, or I thought, that very turbulent times with the high need for government action would lead to much less evaluation. I think what we are seeing is that it is leading to a different form of evaluation, where with that pressing need to act on the edge of chaos, to use the earlier account, that what the evaluators can seek to do is to map what is going on. We heard about that: look at the basis for a future evaluation, look at the timeliness of the response, look at the legality of the ways in which it was being implemented, and begin to develop a framework for the future, I think that earlier Michael and I were talking more about where there was greater uncertainty, but not on the edge of chaos, and that is where emergent approaches to real-time evaluation become quite appropriate.

Probably what we have not talked about is that in those areas where there are high levels of technical certainty and high levels of agreement, there is still a place for traditional classical evaluation. So it does seem to me that it's got an interesting sense of three different approaches that might become appropriate, depending on how far away from that access from that earlier slide. We know as you get towards chaos, actually even doing scenario planning isn't going to help a great deal, but you begin to try to stabilize the future and stabilize the sense of understanding and build evaluation frameworks for the future.

Nidhi Khattri

Coming back to the issue of independence, when you conduct these evaluations and if there are specific recommendations particularly around implementation or design, is there any thought or any issue in your minds about then recusing oneself from doing evaluations down the line of the same projects or programs? Are there any issues with respect to conflict of interest, because if recommendations get adopted, is that sort of stepping into the kitchen some. So your thoughts on that would be helpful.

Stephanie Shipman

Those are great questions. The independence from management is handled the way GAO has for, I was going to say centuries, but it is really just decades done it. We do not prescribe the specific management actions that should be taken, but rather recommend that appropriate actions should be taken to fix this problem. So management then has the responsibility of determining what that action is and putting it into place. We carefully specify the nature of the problem, what is lacking—guidance, procedures, review, whatever—and that it is management's job. That is what we do day in, day out, year after year after year. So we do not get into that problem.

We are reporting at a much more rapid pace than we normally do, and this review with the Recovery Act is a way to allow mid-course corrections, essentially prompting more guidance, better clarification of the requirements, and the like so that OMB can keep building and rebuilding guidance to make those changes. But that is their responsibility, not ours.

How is it used by management? It is so tailored, it is absolutely tailored, all the reporting and the discussions. Again, a lot of real-time briefings for managers at OMB and in the individual federal agencies about what is going on that allows that process to keep going. So what you will see in each of those bimonthly reports is a little update on what happened to the recommendations that were made in the last report and the like.

Structuring with reference to attribution: That is a big one, and then there is the fast action, poor results issue. The agencies already have a variety of processes and procedures in place to do a lot of the efforts. Okay, so with the transportation funding they were encouraged to pick the projects that were top on their list to be funded. We know perfectly well there were projects waiting to be funded that already had the planning and the bids and the proposals, so they were ready to go, shovel ready, that is what that is all about. So you are reducing uncertainty tremendously and allowed to be able to prove the quality there by not having them trying to make up stuff because that is when you are creating tremendous risk. We also have encouraged them to use the same performance measures that they were using before, not create new measures just for the Recovery Act. Use the same measures for transportation, for schooling, for school improvement, for hunger assistance, etc., that you were using before you were making use of that knowledge base.

Attribution: We are not really dealing with results at this point. On the other hand, there are two ways to address this. One is when you are as micro as we are getting with detailing the specific use of funds by state agencies and local agencies, you get out of some of those problems. You do not have to make it up. You are actually documenting. There is enough of the knowledge that states were already planning layoffs and other budget cuts in order to meet the shortfalls of funding from their state resources that when the federal resources come in and fill that gap, you're not creating the same attribution problems that you would have in other settings. So part of it is you get around that, and part of it is we have alerted people from the beginning. They should never have said that they were going to create or save X number of jobs, because there is no way anybody is going to be able to provide a good estimate on that. It was dumb. We said that. What we have tried to do is to improve, provide guidance to, or encourage development of appropriate guidance so that they can do better estimates at the local level of what these particular dollars paid for. Who was employed under those dollars?

Philip Airey

How do we maintain our independence but at the same time engage with management? We are quite clear that we are independent, we report directly to Parliament, but we are always open to informal discussions with the managements of government departments, and indeed we encourage them to approach us informally and chat things through before projects, during projects and after projects. That works well, and everybody understands the rules. They will not use it against us if we get it wrong and come along and criticize them afterwards. I think it is good that we do that and we should continue that way as long as we maintain that informal understanding between us.

I heard a question on how have our reports been used by management? We have had some impact. The primary impact was about accountability and transparency. There have been changes, or changes are now taking place in the way that the Treasury will recruit professional advisors in that sort of situation. Some of the contracted private investment banks in 2008, particularly, were not ideal, but it did have to be done in a crisis situation. We did recommend some changes there, and some changes are now being brought in. We also made a series of recommendations for the Treasury internally about how they organize projects and how they oversee projects like this. Again, this was a crisis, but their project manager techniques are now being put in place on the insurance scheme they are now looking at, unless there is a very, very different project to the early crisis response. So we have had some impact and literally at the margins becausewhat we think they did was pretty good. So we are acting there in the margins.

Somebody asked about how we judge the success or otherwise of individual interventions. I think we realized pretty quickly that we were looking at a program of what were, at first view, a series of interventions of different types at different stages, but we quickly realized this was a program that was being managed by the Treasury. It was all done in response to market changes and risk in the financial sector, and I think, as I said in the presentation, you could pull out one of those schemes, for instance the guarantee of bank borrowing in the markets, where banks are allowed to borrow privately but with a government guarantee. There are measures of that, you can look at interest rates and other things, but you can do that, and you can say, yes, they are heading in this direction, that is good, bad, or indifferent. But ultimately, there certainly are other things that would have an impact on this. As I said, the monetary policy of the UK government or interventions by other UK government departments that are not part of this program and what is happening in other countries and sentiments in global markets [have an impact], and we realized that we could take it only so far.

We could have gone down the route of attempting a complete evaluation. It would be probably too early now, but I think that was way beyond the scope of what we wanted to do at this stage. I am not ruling out that we might do something on those lines in some future date, but I think it is unlikely. I think it is just too complex, and what purpose would it serve? If there was a prospect of being able to look back at what was done, and when it was done, and how it was done and imply that some schemes were perhaps more effective than others and, therefore, if you get in that situation again, do this first and then do that later, if need be, then there might be some value, but that's something that we will keep in our minds as we go along. I think it's a wider issue for when we begin to look at the changes that are now being made to the regulation of banks and to the way that government itself will organize that.

I think somebody asked about the "do-nothing" option and how we actually came to the opinion that do nothing is not an option. I mean it actually came down to the position obviously of banks in an advanced economy and the impact a major failure would have. I think one thing that struck us [was]when we looked at the government's handling of the crisis at Royal Bank of Scotland (RBS), which quickly ran out of money and had to be supported with a series of loans by the Bank of England. What if the government had allowed RBS to collapse? RBS had assets of, I think, two and a half trillion Sterling, which is 'way more than the annual GDP of the UK for start. It is just gigantic, and it had 20 million customers. It was a counterparty to lots of other banks in the UK internationally. And then the modeling of a disorderly failure—there is no way of letting it fail in an orderly way, unfortunately—said it would have been catastrophic, ultimately with a contagion effects in other banks, as well. The UK was looking at a potential breakdown in social order if they had allowed to it happen. So these were some of the most important decisions taken by UK governments since the Second World War, and I am not overplaying that. There really were difficult decisions.

As an audit office, we felt that we needed to be quite clear when we were going into this that the government had to do this; we were not allocating blame as to why they had to do it, and we thought of this, but it was something that needed to be done. I am angry about Tom's point about our reaction to this, given the circumstances, it was quite a scary moment for us. What do we do? This is 'way outside our comfort zone, we look at lots of programs around government in health and defense and all over the place, and they do tend to be at that bottom left-hand corner of the complexity graph, most of them. I think there are well-established benchmarks by which you can measure procurement of defense equipment, health programs, all that sort of stuff, and we have been developing those for many years. This, when it happened, took us out to the edge of the chaos frontier, and we were not prepared. Treasury certainly was not prepared. We had a bit more time to think about it, and what we have come up with is certainly a first stab at it, that mapping report just trying to set up for outreach to back bench members of Parliament, media, the taxpayer, and generally in the UK. A lot of time they were just scratching their heads, saying, "You didn't understand this at all." Why did they not just let these banks fail? Why were they protected? Why was loads of my money put towards protecting these people? We felt we were really on sort of a mission to explain, to bring a bit of transparency, to bring a bit of accountability to it.

Ultimately, as the years go by, we will reach judgments in value for money. The UK government has said publicly that by the time all these schemes are wound down, there will be a return for the taxpayer. We are discussing with them what they mean by return, but we will actually come back around in a year's time and say, well, actually, did they get a return, not necessarily a profit but some sort of return? We will evaluate that. So, it is scary stuff, it is big, important stuff and, it is a pleasure to do.

Ali Khadr

Just to pick up on the questions that Roland raised. On attribution: A very good question, I think, which we always grapple with, and particularly in a complex situation it is incredibly challenging to try to attribute a cause and effect. I think where we are at the moment is that we can say something directionally about what is happening. For example, where finance is being withdrawn and IFC has come in we clearly see there was no alternative. Then we can see that IFC has played a role, it has shown some additionality. We look at the facts on market confidence. I mean, that is a directional thing, and the extent to which the financial sector was stabilized. That has been an issue in a number of countries in Europe and Central Asia, an issue of course in the UK.

On IFC, specifically, you were asking about the findings on past crises and what we had found about which projects have had the most effect. The conclusion I think we have reached is that if we are right in, immediately when a crisis hits, that can have an effect on ongoing operations, and it is not necessarily wise to invest right away, but as soon as you have hit rock bottom, which in this case was very quick, it is relevant and important to invest. In fact, IFC was estimating all sorts of demands for new investments just last December. For example, the potential equity investments in banks were estimated to be about $30 billion. It was a rationale for setting up the new initiatives. Incredible demand was out there. Other international finance institutions have managed to grow their operations in these difficult times. Of course, we will see results down the line, but directionally it seems that there have been some sensible interventions, and that includes IFC.

We are not saying there have not been some successful investments in banks. In Georgia, they have had some very good short-term effects. But what comes out of the analysis is a sense of missed opportunities. The demand has been out there. That was recognized. It was the rationale for going ahead with these initiatives. We see on the ground that first-class clients were needing support from IFC and were not getting it. So that is the overall flavor. In terms of managing the portfolio, it is useful to have a sound portfolio, and that will help IFC in the future. It will help future operations in two or three years, but it is not a response to the crisis, and that is really the fundamental point.

Daniel Crabtree

I would like to come back to one point that was raised, and a point that was addressed also by Stephanie Shipman, on the extent to which one should have precision in the recommendations as to the way ahead, based on evaluation, and the extent to which in our language that takes us into the kitchen and into a more management function that equally undermines independence. Stephanie laid it out very nicely that ideally, one would like to diagnose and say this needs to be addressed and perhaps even outline the prioritized areas that need to be addressed. But I have found that, in reality, it is not that simple, and often one gets a very direct comment from Board members who, after all, we report to, not to say management and other stakeholders, that it is all very well identifying areas that need attention, but, we also need some wisdom on the relative efficacy of measures, different measures, given what we know. That downdraft, I have found that a lot of my colleagues and I struggle with on almost a daily basis: how to be precise, yet without compromising our independence, because if you make a recommendation that is too precise, if you, to put it another way, get into the kitchen, you become compromised, and to what extent can you then evaluate future programs? It is a tough issue and, unfortunately, I do not think I have a response.

Marvin Taylor-Dormond

Every report that we produce contains recommendations, but in this case we have only indicated issues going forward. So we are not recommending anything specifically, very much in line with what Stephanie has mentioned that is a normal practice, as a matter of fact, in GAO. We clearly understood that we could not recommend in a real-time evaluation context.

Mark Sundberg, Manager, IEG

Ranging from the morning ones that were abstract and theoretical to these applied cases, all of which deal with crisis and largely budget transfers or budget support loans in Bank parlance. Of course, much of the evaluation work that we do in the World Bank -- and we have three groups represented here, IFC, World Bank, and MIGA -- is on projects, be it infrastructure, with a long duration period, or a national education curriculum, or down to community level practices. I think the real-time and adaptive evaluation issues are very pertinent here too. The World Bank model of evaluation could—this is too simplistic—hardly be further away from what hass been said here, in the sense that we evaluate projects after closure, so it often has been years after they have been initiated, and then even a lag after their closure. And we use objective-based approaches, so you are confining the questions that are asked. If we move towards an adaptive evaluation model, I think it forces us to really get involved with posing evaluation interventions or analyses at the supervision and at the entry stages of projects to build that into part of the process, which raises questions about independence.

So my question is twofold. One is how would you characterize designing that across these very different sorts of projects from national interventions of long lag periods to very local community orsmall interventions that are perhaps more easily addressed? You have mentioned evaluations, Michael, where impact evaluation is more pertinent at a granular level in that certainty area, I think you used the word simple. But what about across these sorts of projects that we deal with? And secondly, given that there is a very attenuated causal chain, from the Bank that deals with donors, deals with national governments, down to local level government, down to local implementing agencies, how do you build that in across these complex areas of evalution that we deal with?

Keta Ruiz, Senior Operations Officer, IEG

I would like to bring a question pretty much in line with Mark's question for a more specific case. I am working on an evaluation of the Information and Communication Technologies Sector and the support of the World Bank Group on that. This is very *sui generis,* and I would categorize it as complex sector because there is a lot of innovation and technological change. The markets are changing and reacting to these technological changes, and there is the role of the public sector that is also quite different in different countries. So there is a lot of complexity, and one is the role of the World Bank Group and how well the World Bank Group has been supporting the client countries in this sector. There are the infrastructural kinds of projects that we support, but then there is this kind of project or this kind of sector that is very, very rapidly changing. Michael gave one suggestion of a methodology, benchmarking the private sector, for example. I would want to hear a little more about what methodologies could be used for this kind of innovative, rapidly changing sector.

Ismail Arslan

Actually, I am not going to ask questions, I would like to answer some of the questions raised by Mark and other colleagues, particularly on the World Bank side. What we are doing in this evaluation is looking at large design issues rather than results. For example, some of the infrastructure projects are in response to the crisis, designed in such a way that they are responding to the impact of the crisis, creating short-term employment, either in Bangladesh or Ukraine. The other point I would like to make is that this evaluation in the World Bank Group has two stages. In the first phase, we are evaluating the World Bank Group's response to the crisis. For example, the World Bank is investing heavily in economic and sector work; part of our fieldwork is on the timeliness of World Bank's reports on country economic memoranda, or poverty assessments. The second dimension we are looking at is on lending; as my colleagues mentioned, we are working on design issues. Very few loans have closed yet—they are still under implementation. In the second phase we will be looking at more interim results and impacts.

Michael Quinn Patton

Taking on Mark's question about what would be a more real-time role for the Independent Evaluation Group, let us suppose a program design that is more complex-based and emergent, and then look at IEG's role in that, and building on Stephanie's warning, which is regularly ignored by politicians and program designers, about not picking big, hairy, audacious goals that you have no way of meeting, because that is actually more about politics than it is about delivering anything. So those things get set, and then you deal with them. But, one of the recommendations of people dealing with complexity on the management side, the management gurus, the Jim Collins, the David Snowden types, Henry Mintzburg, is that when companies begin a strategic effort or project that they not start with predetermining what the outcomes are, but recognizing that one of the outputs of engagement is outcomes, that you do not begin when you have not engaged. Under conditions of complexity, the goal setting is not done ahead of time, it is done as a part of the engagement, when you know enough.

The International Development Research Centre (IDRC) in Ottawa, did what I think is a fascinating and instructive study of their big five- and ten-year programs, about timing the role of evaluation feedback and reporting, and the question that they ask of their senior people is when is the greatest learning and adaption taking place in programs? Think about that for a moment, some of you who have been on the management side, when is the greatest learning and adaptation? You have done as they do; they spend 18 months to two years coming up with five- and ten-year proposals in big program areas, working with people around the world, and so then they begin it. What they found is that when the rubber actually hits the road, in the first six months of implementation, good projects change all their parameters, because now all of what they assumed, all of what they questioned—Are the resources there? Are the players there? Can we hire people? Do we have office space? Are the partners really going to engage?—all of those things get real, which were in the assumptions column, and they redesign accordingly. That redesign can be well done, it can be badly done, and one of the outcomes of that is typically a closer set of indicators to the real action.

That is a key place for an independent view, because the original program has gone all the way to the Board, been approved by the Board at one time for five years, but in reality it all changes in six months. That is never approved again, and so the need for independent review.

IDRC is quasi-governmental; they are funded primarily by Canada, but they get foundation funds and other kinds of things. They are reviewed by the Treasury Board and by the audit authorities in Canada, and have had to develop procedures that have public accountability around them. So it is a quasi-governmental group.

So the first place where an independent set of eyes actually is needed, that does not happen, is in the big adjustment period, when things get underway: the reasonableness of those adjustments, the reasonableness of new outcomes, the reasonableness of new program designs. What I was saying earlier under the notion of what gets measured, gets done is that currently the performance measurement metric is having people preset outcomes and holding them accountable for those. But if we measure resilience and adaptability under conditions of complexity, we will get resilience and adaptability. If we measure conformity to preset, but unrealistic and not very meaningful outcomes in a narrow mechanistic accountability mode, we will get compliance, with the pretense that those are real and attempting to meet them. So there very much is an important role from an accountability perspective about what that means, and what it dominantly means is in the simple box.

The reaction of evaluators on the whole to complexity is to try to control it, to believe that the way that you deal with complexity is to impose more control. Nothing could be more wrong and more damaging. Under a do-no-harm modality, I think evaluators do a lot of harm by imposing fixed designs, by requiring fixed indicators, by holding people accountable for fixed indicators, by actually interfering with adaptability and resilience because of narrow mechanistic accountability frameworks. So we bear responsibility here.

Independence is also rigid in some cases. We impose rigid models that keep people from adapting, and that part of independence then is assuring the more general public and taxpayers that adaptations that are made are reasonable, that people do have reasons to change what their outcomes are and to adjust. That greatly cries out for an independent set of eyes making judgments about the reasonableness of that, because as Tom pointed out, the danger here is that anything goes. I had a foundation president describe complexity to me as a program officer's wet dream, and it is precisely on this issue. So I think it what it requires is faster, more flexible and different rubrics of what independent accountability means in a real-time, unfolding, complex kind of scenario.

Tom Ling

You asked about how to characterize the kinds of evaluations that we are talking about. I just jotted down a few things. There is moving from doing studies to streams of evaluation, from post to real time; from objective outcomes to contribution stories; from fixed outcomes to emergent outcomes; from detached evaluators to embedded evaluators; and from proving what happened to understanding what happened. I have certainly done evaluations which have been on the right-hand side of that, the latter side. In each of them I have had to take clients with me through that journey, who initially might have been a bit worried about where that might take them, but on reflection, of course, I have never done that for the European Commission and nor have I done it for the NAO, nor indeed for the World Bank, it so happens, who have this public audit function. There is a certain institutional architecture that places a different set of constraints on an evaluation conducted within that framework than for evaluation more widely.

I think there is an issue about how you, how we, think about the arguments that came out earlier this morning and compare that with the efforts by public audit bodies to make sense of an emerging and critical situation. But yet, you know how uncomfortable that might make you feel because of the institutional setting within which you are operating. So it might be quite important for us to think about how we can take the lessons learned and the discussion that we have had, but also understand how that could work within the particular settings of the World Bank, or the NAO, or the GAO, where I believe they have different requirements placed on them than I would as an evaluator working for RAND or a government department.

CONCLUDING REMARKS

Daniela Gressani

This has been very, very interesting. I do not think that I can say thank you fully to both the presenters and the participants, but I certainly have learned a lot and I think everybody else has learned a lot. We got a lot of food for thought, which has direct relevance to things we do and therefore is especially valuable. I have been wondering whether I could abuse my privilege here, not try to summarize the discussion, but to tell you what my three top take-homes will be, and I am sure that different people will have different ones, depending on where they sit and what their key priorities are, but I just thought it might be worth mentioning.

The the first take-home for me would be that real-time evaluation has become the norm among institutions like the World Bank Group, which are large and complex, because of necessity. Michael, I think, referred to the Black Swan phenomenon. I think what it means is that we do need to provide real-time feedback or real-time learning. We have no choice but to be able to deal with it, get organized, do our best.

The second take-home for me is that we need to live with risk, with uncertainty, and with interdependencies. So in my mind, that really requires a lot of clarity about the framework within which we are doing our evaluation. Whether we are in the simple corner, or the chaos corner, or somewhere in between, I think it is important for us to be very clear about that as we launch into evaluation.

The other take-home for me is the fact that I think everybody has mentioned, directly or indirectly, that we cannot just evaluate by objectives. We need some, I think Mark used the word adaptive models of evaluation, something that allows us to avoid mechanistic approaches, that requires that we use good judgment. In order to be able to use good judgment, we need real independence, we need enough resources, and we need something that, I am not sure who, refers to as trust. A constructive, engaged relationship with all of our stakeholders and, first of all, Management and the Board, which allows us to communicate directly and constructively and to trust one another, that we mean what we say and we say what we mean.

Clearly, this is a big challenge. I do not think that we at IEG have ready-made answers, ready-made solutions for delivering on this kind of objective, and as I thank everybody here again, I also would like to get a promise from everybody that this is a first engagement, but certainly not the last, and that we will need to keep learning from one another and we need to keep having a very open mind to learn from our own lessons and mistakes and successes as we go forward.

Marvin-Taylor Dormond

Talking about instant real-time evaluation, I just want to say that using the nomenclature of the NAO, I really got value for money here. That is my immediate evaluation of what we have done during this morning. I think it has been a fascinating discussion, and the two initial presentations beautifully set the stage for what we had in mind, coming from the same context as was developed by both presenters and presenting compelling arguments to move ahead in the area of our real-time evaluation and prospective evaluation.

I think it was really interesting what we heard about prospective evaluation, using either scenarios, as was presented by Tom, or using assumption testing, as has been the practical use that has been introduced in GAO. There are some ideas that we will have to explore more in the future, and I very much agree with Daniela that this is just the initiation of this conversation. I am sure that we will meet again and try to compare notes with what we have been doing. I am really comforted by hearing that what we have done in impact evaluation seems simple, but it was a huge change here, and the way we have done things in IEG is a huge mental model change for everyone, for the Board, for management, and for our own team. It was not easy to start navigating in that direction, but I am comforted by what you have said. It was not an option in practical terms—we had to do something in the midst of this gigantic crisis. But it is clear that it is not an option from the conceptual point of view, either. So there are very strong conceptual arguments as to why we should continue embarking in this direction.

KEYNOTE ADDRESS

INTRODUCTION

Patrick G. Grasso, Management and Evaluation Consultant

Our keynote speaker is Mr. Michael Quinn Patton. Michael spoke earlier today, of course, but let me give you just a little bit of background. Michael was on the Social Science Faculty at the University of Minnesota for some 18 years. For five years he served as Director of the Minnesota Center for Social Research. Most of us who have been active in the evaluation business for a long time have known Michael very well, in many capacities, one of which was president of the American Evaluation Association and, as the author of quite a number of books, and coauthor of many other books and articles. He is probably one of the most widely-read and recognizable names in the evaluation business. I very rarely see any major book on the topic of evaluation in which Michael Quinn Patton's name is not somewhere in the references. So, it is with a great deal of appreciation for his coming to visit with us today, and a great deal of anticipation at his comments, that I would like to welcome Michael to please come up and give us our keynote speech.

COMPLEXITY THEORY AND EVALUATION

Michael Quinn Patton

I think this is a tremendously important meeting and discussion to bring together people both within IEG and some of the other parts of the Bank, and the people who have been resources from outside in other organizations that are struggling with these issues of how to adapt evaluation to both our changing times and our changing understandings of our times, and that is really where I would like to begin, with how we understand what is going on and the importance of spending time on that.

For the distinguished folks who have joined us at lunch, let me quickly review what we have been doing today in talking about real-time and prospective evaluation. Basically, we have been looking at the implications of things like the global financial crisis for engaging in evaluation sooner rather than later, which is being called real-time evaluation, getting feedback about how these interventions under conditions of crisis are actually happening, and doing that in a way that provides both some public sense of accountability and internal guidance about improving those responses to crisis situations. We began the morning with some overall conceptualizations of the problem and then heard from people who are actually doing this kind of work both outside and inside the Bank. What I want to do is push that discussion—being an author and researcher about evaluation, and myself trying to keep up with these new directions and writing about them—to share with you what I see going on here and, in so doing, to be provocative about what the issues are and some of the opportunities to respond.

Interpretive frameworks

Let me begin with the whole notion of the importance of the interpretive frameworks or the interpretive mindsets that we have for whatever arena that we are engaged in. There has been a huge amount of work in the social sciences in the last decade about how we program ourselves, through socialization and within our culture and within our organizations, to see the world in certain ways. And within the world of interventions, the dominant way that we have come to see things is in linear, mechanistic ways, constituted and represented by things like a logical framework, logic models, and the notion that interventions are aimed at pilot testing something that's going to be replicated and taken to scale throughout the world. The notion that evaluation is about testing models is fairly dominant, and that's a mindset. It is an interpretive framework that constrains, and issues of accountability, independence, and performance measurement and results are all affected by and reside within that framework of things.

What we have been discussing today are the implications of an alternative framework represented in simple language by the term complexity or complex-adaptive systems, where things are highly interactive, rapidly changing, not predictable, not controllable, nonlinear, and where our knowledge base about what to do and the agreement about what to do is fairly minimal. So high degrees of uncertainty, high degrees of disagreement about what to do, and situations where, in fact, any intervention within a system creates actions and reactions that are non-predictable, that are iterative, that come back, that go forward in unpredictable ways, and indeed one of the graphic images of the morning was a knot all tangled up so that you could hardly tell where the ends were. What we typically do as evaluators is to think our task is to unravel that knot and find the straight lines rather than deal with the knot, and to even think that we are not changing the situation by unraveling it and trying to make it straight, instead of looking at the intrinsic and sometimes helpful characteristics of the knot itself, and dealing with the knot. So these metaphors are part of what we are going to play with.

Where I would like to begin with is some intriguing research done by two management organizational development scholars at the University of Michigan -- Kathleen Sutcliffe and Klaus Weber -- in a 2003 *Harvard Business Review Report*[13], in which they compared two sets of high-functioning organizations, each of which was going through major strategic processes and strategic planning processes. One group of organizations decided that the way to get better at what they were doing was to measure their performance more precisely and to use the best practices around performance measurement, and they set out to do that and put resources into getting better data, larger sample sizes, [and] understand-

[13] Kathleen M. Sutcliffe and Klaus Weber, "The High Cost of Accurate Knowledge," *Harvard Business Review,* 2003.

ing the knowledge base of their fields, across diverse industries, better. The other group, going through strategic thinking and looking at their situation, determined that they were probably in a highly dynamic environment, alluding themselves by thinking that they could get precise measurements of moving targets and that what they needed to do was to spend more time at senior levels, making sense of the data they already had and that they [had] access to and that was coming in, indeed, in real time. That they needed to spend more time interpreting and less time worrying about the precision of the data because it is an imprecise world. They followed these two sets of companies over time to look at how each was affected by their performance, and what they found is that the companies that define the situation as understanding and responding to their environments did better over time than the companies that thought the issue was more precise measurement of their environments. The title of that article is "The High Cost of Accuracy," which has to do with how we define the situation.

So part of what you are faced with at every level in the World Bank is how you are defining a situation, and much of what has gotten defined in the situation. In IEG, what we heard constantly this morning is that it is maintaining independence, maintaining accountability, being able to specify attributions. I am going to suggest to you that those are old paradigm concepts, they are mechanical, they are largely outdated, and they are interpretive mindsets that actually become barriers to dealing with the complex realities of a rapidly changing world.

Let me give you an example of an interpretive mindset and why we need to dialogue about how different people can take the same data and reach different conclusions. There is a story about a man who was very, very ill, and after some time in the hospital, he was getting well, and as he was about to leave the hospital his wife met with his doctor, and she said, "Doctor, tell me the truth, what's the real story here?" The doctor said, "Well, your husband's been really, really sick, but if you take really good care of him, give him the kind of food he wants, loving, and give him all the sex he wants, he'll really be okay." She said, "Well, thank you, and I appreciate your being frank with me." So she came out of the doctor's office, and they started walking out of the hospital, he said,

"So, what did the doctor say?" She said, "He said you're going to die."

Complexity and evaluation

Now part of what interpretive mindsets mean is that the kind of data that come in and the framework about data under classic and traditional evaluations is finding definitive answers to did it work? In complexity situations, we do not get those kinds of data. We get patterns, we get feedback, we get possibilities. We are dealing with moving targets in a rapidly changing world.

I encountered this challenge to my own mindset some years ago, when I was doing a local evaluation of a leadership program in northern Minnesota by the Blandin Community Foundation that was trying to train rural leaders throughout the state of Minnesota. They brought them together in an intensive retreat environment that they had never experienced before, gave them [training in] communications, strategic planning, dealing with indicators, [and] how to do community organizing and sent them back in their communities. I have the evaluation contract to do two and a half years of formative evaluation followed by two and a half years of summative evaluation. Classic, probably the most classic form of evaluation contract, and they were a great group to work with. They were open to formative feedback. They kept changing their program. We followed people up, found out what was working for them and not working for them, they adapted the program curriculum; they were very open to feedback.

On a cold Minnesota morning in February, I met with them after two and a half years and said, "You folks have been a great group to work with, you've been open to feedback, you've made changes, you've really adapted, but now we're moving into the summative period, where we have to decide if the model that has been developed works, and so you can't make any more changes in the program, because if you keep changing, we can't answer the question of did it work. It's got to be stable, standardized, fixed, and that's now the challenge. So change is done, next two and a half years, everybody gets the same intervention and then we'll follow up, and see what's happened to participants, what they're doing in their communities, what kind of differences they're making, how they're communities view [them]."

The director of their program looked at me and he said, "But we don't want to stop changing the program." I said, "No, I understand you've been really good about changing the program, but we're now doing what's called summative evaluation, and that means you can't keep changing the program, the formative piece is over. The Board has contracted me to do a summative evaluation to answer the question does it work? There are a lot of people watching what you're doing. People want to know if they should emulate this model. That means summative evaluation." He said, "No, no, no, no, you don't understand. We understand that we can't keep the program the same, we need to keep changing the program because the world around us is changing." Then he looked at me, fairly hostilely, and he said, "Formative evaluation, summative evaluation, is that all you people have?"

Well, in truth, those have been the dominant paradigms, with an accountability version of summative evaluation, which is a lot of what IEG does. Quite taken back, I said, "Well, I suppose, if you really wanted to, you know, we'd have to renegotiate the contract, but you know if you really wanted to, we could try doing developmental evaluation." And they said, "What's that?" I said, "That's where you keep developing and adapting." And they said, "That's what we want to do. How do we do that?" I said, "Well, we'll have to figure that out. I'll get back to you on that."

It is important to distinguish here that it is not ongoing formative evaluation. The purpose of formative evaluation is to work out the bugs of a model and stabilize the model so that it can be put to a formal test of whether or not it works. Ongoing adaptation is a different animal, and what they understood was the technology was going to affect local leadership programs, and mobile phones were coming in, computers were just coming in, this is the beginning of the Internet age, the ebbs and flows of the economy, changes in regionalization, migration patterns were going to affect what they were doing. They wanted to get more young people involved. They wanted to take this to Native American communities. But they never expected to have a fixed model, and that the role of an evaluation under those conditions, of ongoing adaptation to change, sometimes rapid change, sometimes slower change, but ongoing adaptation to change, is a different animal.

Complexity, the way I have defined it this morning, where we do not know what to do, and there is not an agreement about what to do, means that there is going to be ongoing adaptation, and we are not going to get fixed models to replicate. It also means that the indicators themselves may be emergent and [may] change. This is controversial, this whole issue about when do you have indicators, when do you have predetermined goals, [and] accountability against predetermined goals. But in looking for examples of where people have dealt with complexity and how they have come up against it, it is intriguing.

Performance indicators and time frames

Let me remind you of this, because this is something I suspect all of you will remember to some extent, although you may not have interpreted it quite the way I am going to, and I invite you to go back and check the record and see if my interpretive mindset meshes with yours. But when Alan Greenspan retired in 2005 after 20 years as chair of the Federal Reserve Board, he was going to give his final benediction speech at the Annual Meeting of the World Central Bankers in Jackson Hole, Wyoming, which was basically the world's assembled central bankers and economists. This was his chance to tell the world and that group of people the most important message he had to leave with them about the future of the economy, the global economy, and how to manage it. He could have talked about anything. What did he chose to talk about? You can go back, you just Google Alan Greenspan's 2005 Jackson Hole, Wyoming, speech and it will come up. It was fairly short, and what he chose to talk about was warning the central bankers not to pick indicators and goals as targets. He said, do not do it. For 20 years, Congress hassled him to set targets for inflation, targets for interest rates, and what he said was, any time you pick any singular targets, no matter how many and what subset, you will distort all the other indicators by trying to meet those targets.

What does the Central Bank do? What does the Federal Reserve do? As you know, they have got staff all over the country, they have unlimited resources essentially in terms of data collection and super computers and all that, they monitor all kinds of things, and then once every three months they all come together and they argue about the data. What's going on? There's been a crisis in Mexico. There's a crisis in Thailand. Something's going on in China. There's a more or less global financial crisis. What do we do? How do we make sense of it? They dialogue about that, they have very few, as you know, policy actions that they can take, but they are constantly looking at these moving targets, constantly looking at where bubbles and depressions are occurring and managing this interactive system.

Greenspan's book, *The Age of Turbulence*[14], which characterizes complexity, came out after he retired, and was written before the most recent global financial crisis. What he acknowledged was that there is no model of how the global economy works, and there will never be a model of how the global economy works, and the models we have, have not gotten any better in the last 50 years in predicting the future. That is the definition of complexity. Complexity actually emerged out of meteorology, in studying hurricanes, trying to study weather patterns, and this allows us to think about issues like attribution and causality and management in a parallel metaphoric term. I just spent some time with a group of meteorologists about their work in trying to inform the public of crises and doing warnings. The big picture of meteorology and their forecasts long range over five year periods, not unlike global economic forecasts, is simply that the weather is going to get more and more turbulent, that there are going to be greater variations than there have been in the past, that there will be intense micro weather systems within larger macro systems of change, and that old patterns are not going to be the new patterns. That is not much of a prediction. It is almost identical to what we can predict about the global financial community. It would be absurd, I would suggest to you, on the face of it, to ask meteorologists to tell us what is going to be the precise nature of the weather patterns five years from now, but that is precisely what we ask people running big projects to do. What are the outcomes you are gong to accomplish in five years, in a turbulent economic, meteorological, social, and global context?

However, when the weather gets close they have two-week forecasts, they have one month forecasts, they have one week forecasts and—no surprise—their best forecasts are their eight hour forecasts, which are highly accurate, and are used by people who need to know what is going to happen in the next eight hours: the people who clear snow from the roads in Minnesota and spread salt, the principals who have to decide whether or not to keep schools open, community organizations that have to decide whether or not to hold their events. That is enormously useful, and they get feedback from those people in real time about whether or not their forecast was right within eight hours, because it has big implications. That is real-time feedback, and the quality and speed of those forecasts are getting better, but will never be perfect.

So a part of what we talked about is what kind of performance indicators are appropriate within what kind of time frame. The irony, and you heard this from Stephanie in her presentation about getting into the micro details, as well as in my experience with doing real-time evaluations, is that interestingly the attribution problems actually get fewer, the shorter the time periods because you can connect the dots more easily

[14] Alan Greenspan, *The Age of Turbulance: Adventures in a New World,* New York: The Penguin Press, 2007.

between an action and reaction. It is with long, impact-laden time periods that it is hard to do attribution. It is not hard to do attribution in a short time frame, where there is a very direct and observable action-to-reaction connection, so that the attribution picture actually changes under those conditions when you are looking at how programs that are emergent are actually doing what they are doing. Part of what emerges from these kinds of retrospective evaluations is the question of what we learn from them. Your job here is not to manage those adaptations, but to assure that the changes that are made in real time by managers and by programs in whatever area the Bank activity is going on are well-reasoned, that they are justified, that they are based upon data, and that can be done in real time as you look at how those activities are unfolding. We had examples of that being done in other arenas this morning, and in the work that the Bank has started doing in that.

Learning lessons

But we are also attempting to learn lessons about doing that, and one of the challenges in deciding what we take away from real-time evaluations that has any future use is that whole challenge of learning lessons. One of the things we know often happens is that people take lessons from some event and then end up fighting the last war because future conditions have changed, but they are now trying to avoid the mistakes they made last time, and in fact creating new mistakes because they are not paying attention to how the unfolding world is different. And we just had a wonderful example of that, that if you will indulge me, I will use.

I realize that in an international organization many of you may not pay attention to American television, but how many of you have been paying any attention at all to the late-night talk show hosts fiasco going on? . The full story is that in 1992, when Johnny Carson, the most popular late-night host of all time, retired, there was a big fiasco in picking his successor, between Jay Leno and David Letterman. The network screwed it up. It was very political, very controversial, and so the lesson they took from that was next time plan ahead the succession. So in 2005, they went to Jay Leno, who had the top ratings at *The Tonight Show,* and said, "We want to plan the succession because we learned last time not to wait till the last moment to do this and do it on the fly. We're going to plan for you to retire in 2010, because what we've learned is you've got to plan ahead. We don't quite know yet what we're going to do with you, but you're going to retire and Conan O'Brien is going to replace you." Jay went along with that, assured that things would work out, and sure enough they planned the work and worked the plan.

So, 2010 came, and they told Jay Leno he was going to move to 10 o'clock and that Conan O'Brien would take his position, because they were working their plan. They ignored the fact that Jay Leno had the highest ratings in history and had surpassed Johnny Carson's ratings in the meantime, that this was a very high-risk proposition. They were working the plan, because what they had learned last time was plan your work and work your plan, and do not get distracted by any data. But there was a lot of new data, and, in fact, the experiment was a colossal failure, moving Jay to 10 o'clock and bringing Conan O'Brien in, and so the soap opera of the last few weeks, for anybody paying attention to the entertainment news, has been all this personality stuff, and hurt feelings, and Jeff Zucker the CEO of [the National Broadcasting Coporation] NBC looking bad.

If you read the business press side of this story, which does not get the front page, they are all applauding Jeff Zucker for doing real-time evaluation because Jay's ratings were instantaneously in the tank with no indicators they were going to get better. Conan O'Brien's ratings were in the cellar with no indications they were going to get better, and so they did not set a target for what the ratings had to be. They vaguely said, "We know they're going to start out low and we're going to look for their beginning to attract some people and adapting to those new time slots." At the end of four months, the argument now is whether that was too soon or too late. They did not see any changes in the data, and given the consumer responses they were getting from focus groups and from surveys of people, they did not expect that to change, so they pulled the plug. Now Zucker has been hugely criticized because it has all been about the personalities, and Jay's feelings and Conan O'Brien's feelings, and contracts. But, in fact, he followed the data, and their new lesson is to follow the data in real time. Do not make big, five-year ahead decisions, and stick to them come hell or high water. Now part of the challenge of using real-time data is who is going to act on it, and the politics of action. This gets us to the World Bank Board and to senior managers.

I am trying to pull in some different metaphors and analogies here for you to think about as you think about your own arenas of work. So what I want to invite you to do is not immediately dismiss these as not relevant, but think about what you can learn from these kinds of analogies.

In 2005, at the last International Evaluation Conference (we meet every 10 years as an international evaluation community from the professional associations), the American and Canadian co-hosts gave the first-ever international evaluation award for speaking truth to power to Sir General Romèo Dallaire for his work during the Rwanda genocide. He was the Canadian General in charge of peacekeeping forces, but the evaluation side of that story is that the only lever Dallaire had was real-time reports on what was going on in Rwanda, and that is what he did. He filed daily reports about the numbers of deaths, who was killing whom, what the movements were. His troops were basically on-the-ground reporters of what was happening, and he sent those reports up through channels, and they were ignored. You may know the story of General Dallaire, who came back from Rwanda with huge guilt about not having been able to stop the genocide had a nervous breakdown, was found drunk underneath a park bench in Montreal, has gone through an amazing rehabilitation, and now is dedicating himself to stopping genocide in places like Darfur and other parts of the world.

The lesson that he has taken from that, that he talked about at this international convention, was that he made the mistake of playing a good soldier and only sending those reports through channels, and not looking for other ways to draw the world's attention to what was going on and not dealing with the politics of the situation. One of the challenges to real-time evaluation is the political capacity of large organizations to adapt in their decision making to real-time contingencies.

I was involved in a federal department that I will not name, because it is confidential internal work that I was doing, but in anticipation of the change of administration a year ago, they brought me in to help them design a very rapid reconnaissance of the major issues that the new Secretary of that department would face. And we put together a real-time methodology to look at what was going on. We looked at the evaluation data, the management information system. We interviewed people in the field. The whole thing was done in three months' time reaching out and asking the question, "What does a new Secretary in a new administration need to know about this department?" It was one of the fastest and, I think, best pieces of work that we had ever done.

The group decided that they needed to narrow it down, and they identified five really high priorities, areas that needed rapid response and immediate attention, that had high value and, if not attended to, represented dangers for our country. We got all the methodology right, we got the feedback right, we figured out how to reduce this to a communicable form, but what had not changed in this department was the approval process for getting something to the Secretary, which takes months, and it started going through that approval process. Some of the people carried over from the past administration had a vested interest in not seeing those findings passed on to a new administration. They did not suppress them, they did not sensor them, they did what bureaucrats are good at: they sat on them; they asked questions about them; they sent them back for revision.

I was faced with whether or not to become a whistleblower, because nine months after the change in administration, I learned that these findings had not yet gotten to the new Secretary. Everything was timed to be real-time evaluation, high priority stuff, but the political process did not allow that to happen. What did happen eventually was that people within the department took it upon themselves to leak it, but the timeliness had been severely impaired. So it is not just enough to do real-time evaluation. We have to look at real-time decision making. We have to look at the way in which we are organized at every level to engage with real-time data. It challenges what we have learned about every aspect of things.

Five methodological provocations

Let me very quickly, because we want some time to interact here, suggest five provocative methodological issues. I have been talking sort of conceptually and politically about these issues. I am going to do these very quickly. Some of this is talked about in the paper that I wrote, more of it is talked about in the book that I have coming out in June, but this will at least give you a flavor moving from conceptual stuff to methodological stuff, which mirrors the morning.

One of the challenges of dealing seriously with complex nonlinear dynamics is something that is sacrosanct in evaluation and it is how we understand baselines and outcomes, which are mirror images of each other. Where'd we start? Where are we trying to get to? Then, evaluation basically measures where did we end up? Under conditions of complexity both baselines and outcomes become dynamical and emergent and changeable and unfixed. Now the very fact that that can occur, and in good organizations ought to occur and does occur, increases the importance of having independent ways of verifying that those changes are appropriate and valid.

What do we mean by dynamical baselines? Nothing is more sacred in evaluation than that you have a solid starting point, against which you measure everything else. But let me give you a program level example of this and extrapolate it to a country and international level of it very quickly. I work at all levels, and one of the places that I do a lot of work is community-based, anti-poverty programs and interventions with people in poverty around employment programs and mental health programs and housing programs. They do intake of clients who come into the system and find out what their job status is, what their drug abuse status is, what their family status is, what the history of their family is. What we have learned is that all those people have learned to lie systematically in order to be eligible for the program. They know what they are supposed to say, they know what the eligibility requirements are, and that baseline intake data is absolutely fabricated. It takes six to nine months for a program to build a relationship with those clients before they actually know the realities of what their situation was when they entered the program.

Under current evaluation norms, it would be considered both invalid and unethical to go back and change those baselines. But, in fact, people have entered those programs with much more severe conditions than were originally expected, and that affects the comparison to the outcomes. In fact, in many cases, given the static nature of the baselines, people looked like they got worse during the program, and when experimental designs are done, neither the control group's baseline has changed nor the treatment group's baseline has changed. At the country level, I have talked to a lot of folks and been involved with projects nationally where it is only after you have engaged for about six months that you find out that a lot of the baseline statistics about the project were made up, that the data that was supposed to be there was not real and was not very good. You find out what is really going on in the dynamics between various departments, and most of the baseline assumptions have to be revisited and updated, if it is a good project. That is a dynamic baseline.

The same thing happens with the targets. When you learn more about what is going on and you change what you think you actually can accomplish and under conditions of uncertainty, that is appropriate. Under conditions of high certainty, when those targets are meaningful, where there is a knowledge base to set them, it is appropriate to hold people accountable for them. Part of the very meaning of complexity is that we do not know enough to set targets because we do not know how to produce the outcomes.

That is what it means to be in a complex environment, so it makes no sense to set definitive targets when you do not know how to get to them. You need moving targets, updated targets, but you need to do those with authenticity and validity.

A second issue that becomes very big in complexity is about unanticipated consequences and side effects. Virtually all log frames give token attention to unanticipated consequences and side effects and say they are important and we ought to pay attention to them, but I think it is one of evaluation's biggest dirty little secrets that we do almost none of that in real ways. It is just not authentic. Performance measurement, measuring whether or not goals are attained, is so dominant that all the resources and the designs go into that. The only way to pick up unanticipated consequences is open-ended field work, where you go out to see what happened that you did not even think about might have happened. It is the only way to do it. It is not budgeted. It is not included in evaluation designs. We give the most token kind of attention to unanticipated consequences and side effects. What we know is that in complex nonlinear dynamics, those things are certain to be there, they are going to be important, and they are often more important than the anticipated and targeted outcomes. It means that evaluations, at any stage they are done, have to take seriously the fact that we do not know all of what is going to happen, and we need ways in real time to turn up what is emerging. Stuff is emerging, and it is important stuff, and we were not able to think about it.

We often do not know the consequences. One of the best examples of that on the positive side that I have heard, and you probably have heard this but it made a big impression on me, is that when 9/11 hit, and the attack on the World Trade Center occurred, the world's financial system was virtually unaffected. I remember not long after that, hearing an interview with Alice Rivlin, asking her why they had targeted the World Trade Center and why there had been virtually no ripple effects in the actual financial system. Alice Rivlin's response was because of Y2K. We had just been doing a decade retrospection on Y2K, which became a big, speaking of late-night talk shows, joke. All of that work went into Y2K, the thing that was anticipated that never happened, millions, billions of dollars going in, but the effect of that was to make all the systems redundant, to go through scenarios of backup and what would happen if something happened, to decentralize databases, to run fire drills about what would happen if our systems went down, and 9/11 was the real Y2K, an unanticipated consequence of what went on. So what we learned about these things are the system interconnections over time about a globally interrelated system. The things that we thought were over may, in fact, reappear in other forms, and we need to understand both the implementation and interaction equivalent to those.

How we go about doing this work affects what is done. I talked this morning about the mantra in performance measurement that what gets measured gets done. If we focus all our attention on measuring preset outcomes against preset baselines, that is what we will get. If we measure resiliency, adaptability, and the extent to which people are updating their understandings of situations and setting new, appropriate targets and adapting to that, [then] that is what we will get. Our current system is largely aimed at creating static, mechanistic implementation of programs and initiatives. If we want programs to be able to deal with complexity and adapt, we need to have adaptive evaluation systems because what gets measured gets done.

The fourth point around methods is that under complex, dynamic systems, the findings, require interpretation and dialogue. They are not going to be definitive. They are not going to be black and white. We are not going to be able to say it worked or it did not work. We are going to be able to describe interdependent factors and the relationships among those factors and the ways in which they move together. The discussions are going to be much more like the Federal Reserve discussions, when they look at the data and try to figure out what is happening now, and they decide in six week increments, "Well, that is happening over there, we need to really pay attention to that for the next six weeks and see what happens, and monitor that and pay less attention over here." That is an evaluation process that is adaptive and responsive, and still builds in accountability and can be done independently.

Finally, under complex, nonlinear dynamics and the realities of complexity, there can be no methodological gold standard. The language of the gold standard has done great harm to our capacity as evaluators to respond flexibly to appropriate evaluation designs under different conditions. The very notion that there is such a thing as a gold standard design—which then makes people want to meet it and creates incentives to have that design regardless of whether they are appropriate or not—creates disincentives for new cutting edge approaches.

The "Platinum Standard": methodological appropriateness

The real platinum standard is methodological appropriateness: adapting the evaluation approach to the degree of complexity and the nature of complexity that we face. And that seems to me to be the overarching theme of the morning, both conceptually and in the examples that we heard—that evaluation has to be done in different ways and has different dynamics under different conditions. Complexity represents a different condition, chaos represents a different condition, and the appropriate methods for those conditions are not going to be the traditional evaluation methods that have been more mechanistic and static.

With that, let me stop, and invite both questions and comments from any of you about your own takeaways from the morning. Daniela very beautifully closed out our morning session with her takeaways. I would invite any of you to share with the Board and senior managers here your takeaways and disagreements with anything that I have said.

DISCUSSION

Ali Khadr

Thank you, Michael, I thought that was great. I just had a couple of questions or observations based on a couple of the things you have said. One observation strikes me that evaluative wisdom or the view that evaluation can give almost by definition becomes very time sensitive. Let me give you a slightly different take on the Jay Leno and Conan O'Brien thing, which I heard from a man sitting next to me on a recent flight back from Cincinnati. The person said that the problem is exactly what you said about Johnny Carson and how Jay Leno replaced him, but it had taken a long time for the transition, and what he said was that was real guts. The executives of that time had a vision, they knew their vision was right, and they stuck to it. Nowadays what happens is you observe that ratings are not adjusting immediately, and so you give up on your entire vision, and you adapt. Now you said it is a good thing, but this poor guy seems to think it was a really bad thing.

Michael Quinn Patton

Very quickly, what got left out in the story is that in 1993 affiliates were not powerful. The pushback here was not actually the ratings, it was the affiliates' pushback, the people whose news programs were being hurt, who were threatening real action and saying, "We're monitoring the situation in real time, and we're going to stop carrying your shows if you continue another month." So there was a real threat. They had real-time feedback: "Do something now or we're out of here." That was not the previous condition, so the world had changed in terms of the power dynamics between the affiliates and NBC during that time.

Ali Khadr

Very good point. Second point, very quickly, is just on the issue of updating baselines. Again, you portrayed it very much as a virtue, as an issue of responding in real time and so on, and it is a great way to look at it. Think of it another way, though, which is that if I can say to somebody, "Well, you know, I'm going to have to update my baseline, that's a virtue, right? I am going to be held accountable, but only for updated baselines." And guess what, I can influence whether, at the country level, information gets gathered or not gathered, and I can keep updating my baseline and say, "Well, sorry, but we're not getting good information, and so we have to keep updating this." I can work my way out of accountability, and I think that is part of the explanation as to why the incentive system at country level is so slow on the results agenda. The point is that good data are not technologically difficult to gather, yet it does not happen or it is not happening fast enough. I just sometimes wonder why and what the sort of incentive framework is that gives a result like that, where this is so technologically simple to know what is going on out there.

Christine Wallich, Director, IEG

Michael actually put a finger on some of the challenges that we are facing, not just in evaluation but in development. On not being rigid on original goals: We have to keep our eye on something, and development takes time. You have to have some guidance so you do not lose your original purpose. But we design projects with the best intentions, with the best predictions at the time, and then it takes about four to seven years to implement, and sometimes it could go up to ten years in fragile states. So many things change during that time, and original goals often are not totally valid or achievable, or you can overshoot [or] undershoot. Anything could happen. But our current methodology for IEG, for the World Bank is sticking to original objectives, even if you restructure a project, because management is encouraging this responsiveness [and] adaptiveness, which means you have to restructure as you go along. But everybody knows that you are going to be evaluated by the original objectives, even if restructuring was done on valid grounds. So that is something for us to think about collectively.

Second, I missed the morning discussion, but real-time evaluation implies that things are happening before you really get to your goal. It means that when you do real-time evaluation, you are looking at intermediate signs and early warning signals and more input-output measures than what is real impact. We have to keep choosing what to look at, and deciding whether it provides good predictive signs for the long term, for what is going to happen. It is very challenging to find things that are really predictive. It sort of changes the total evaluation concept because our evaluation concept is that you have to look at the impact, and if you do real-time evaluation, it means you do not wait until the impact. You look at intermediate things that are happening. So we have to think about that.

Also, I would like to think collectively about the roles of independent evaluation versus management. Staff learning is important, and the whole objective of this seems to be learning. We all have to learn and adapt very quickly. So what should independent evaluation bring? What should management be doing about their own self-evaluation and adaptive implementation? We have impact evaluations, a little bit [of a] different animal, but they could be done at different stages. You can do it in the design stage, you can do it ex-post, it could be done in different stages that will teach you different things. So all these are useful things for us to consider.

Roland Michelitsch

Hello, Michael, a couple of observations. On some points, I would completely agree that you don't want to spend too much time getting more and more and more data. You need to also take the time to evaluate that. But in our organizations it is the case that even basic information is often still lacking. So I just want to make a plea to not use that as an excuse to not collect basic information, which is a huge problem, and I think too many decisions actually are being made with lack of data rather than having too much data, at least when it comes to development results.

Secondly, contrary to what you are saying, I still think it is a very good exercise to come up with clear projections, not only for five years out but also for the next year, and the year after that, and so on, because only then can you actually track whether this project is on track or not, and you actually can prioritize where you are going to put your resources: checking why is this off track, can we actually put it back on track, and then putting it back. In a sense, I do think that in IFC we have something like real-time monitoring or evaluation, or whatever you call it, with the Development Outcome Tracking System. There we do require people to make five-year projections and also annual targets, but what we then do is, in terms of evaluating the performance, we try to use absolute benchmarks rather than only the objective-based benchmarks, so that you can say, "Well this objective I overshot, this objective I undershot, but overall, using objective benchmarks, in economic terms does it generate above 10 percent return, does it meet the environmental standards, and so on?" So you need to have a combination of the two of them, and I would really shy away from a message going out of here that people should not be setting clear targets and objectives up front.

Hans-Martin Boehmer

I want to take advantage of the fact that we have some CODE Board members here as well. Two years ago President Zoellick had a long discussion with the Board on his strategic vision and the quintessential diagnosis that he had was that the task of the World Bank Group is to help solve interconnected, complex, dynamic problems. It described exactly the world that Michael was just describing about complex systems and limited ability to really predict what is going to happen, and he gave lots of examples of why that is the case from the food crisis and so on. The Board still has discussions with the president on post-crisis strategy, and it strikes me as if the role of IEG is in some sense an integral part of the vision that emerges from that. If it is, in fact, one where the purpose of the organization is to deal with this complexity, and to accept that there is a world of unpredictability or a dynamical world, as we heard this morning, as opposed to dynamic, where everything goes perhaps in the same direction, then that would imply quite a different role for IEG. I would be interested to hear what kind of evaluation function is commensurate with that direction, whether that is something to be thinking about, and whether that is something that perhaps should be discussed.

Giovanni Majnoni, Executive Director, World Bank

I would like to thank IEG and Michael Patton for the extremely interesting conversation over lunch. I would like to just mention two little points which are basically my take, and the second is also a question for Michael.

The first is that this complexity is in a way something that has always been built into social sciences, so in the way we study something our very understanding of what happens affects the outside reality. Globalization is nothing but the magnification of this, so that to a certain extent we can expect what we do—our policies actually are affecting the way the process presents itself. I somewhat disagree on the Greenspan comments, as a former central banker, because there is a good clause which says that whatever you try to target by being targeted changes over time. So I would give more credit to central bankers.

This brings me to the second point, which concerns baselines. We live in a world where baselines are often non-existent, as you mentioned, and therefore I find [it] hard to swallow that the little we have may disappear for a greater good. So the way I think, and this is a question, is that baselines should not disappear, but maybe we need a range of baselines, an upper end, a lower end, and this thing should move over time. This brings me back to the central bankers, who typically, when they project the monitoring aggregates, all have this range, which moves and every year is adjusted. So in that sense is the above conclusion, to judge from your remarks, widely out of line?

Konstantin Huber, Executive Director, World Bank

Thank you very much for this interesting lecture. I have the impression that probably the world is even more complex, insofar as there are highly complex parts and other less complex parts, and we are still to find out what is what. Now dealing with the financial crisis, of course, everything has been turned upside down and things are developing extremely fast. So I very much appreciate your points in this context. On the other side of the development context, the traditional development role of the Bank, we deal with an environment which is at times not developing very quickly. I am a development practitioner, and I have spent many years in developing countries. And looking at the situations there, I am sometimes overly surprised that things are still the same, and they still have to tackle the same issues and the same problems. In that context, if we do not have a good baseline, we do not try to understand the initial situation, we never get anywhere. So I think it is probably at both ends—yes, trying to be adaptive and trying to grasp what is changing, but also going down to the baseline and getting the data. And I completely agree data are not there, but they are not that easy to get. It is a matter of the government, and it is also a question of telling the practitioners and operations people to do it, because they want to maintain the flexibility without living with baselines.

Stoyan Tenev

Three comments. One, the most important takeaway from this entire three-quarters of the day session, is that we must adapt, I would not say to complexity, because we have been used to dealing with complexity, that is what we do in evaluation, we deal with complexity about what you call fast changing or dynamical conditions. It is not only complexity, but that events are moving very fast, and we should adapt to these changing conditions, and when I say "we" I believe that it is not only the evaluator that should adapt and change, it is also the users. And I think that it is very appropriate that here we have two important users of the work that we do in CODE and management.

It is not only producing a more appropriate type of work, but it is also having the right audience to be able to use these more appropriate products.

My second point is that I very much appreciate what you have said about with the metaphor of weather forecasts. This just reaffirms what we have been saying in the context of our crisis management work: that you went in there to have short-term results, and in that context you should be able to predict better what your results were supposed to be. It is more difficult to predict what is going to happen with an intervention five years ahead on the road; you should be able to predict better what the short-term results of your intervention will be, meaning that you should be able to put up an appropriate results framework for what you are doing to respond to the crisis.

Finally, I very much sympathize with what you are saying about dynamic baselines. We have a very specific case here. We have been trying to engage in an evaluation of [the] decentralization process in IFC, but the fact of the matter is that over the last three years IFC has been moving so rapidly and constantly changing the conditions in terms of decentralization and organization changes and so on that we cannot find that baseline. What is it that we are going to evaluate because right now we are in the middle of a substantive change again? My question to Michael is how do you evaluate in these circumstances, because the essence of evaluation is comparing against something? One thing is to measure, that is the first step of evaluation; the other very important task is to compare it so that you can judge. How are you going to judge in this change of circumstances?

Michael Quinn Patton

As predicted, this would be provocative to raise questions about the sacrosanct baselines in evaluation and emergent goals. Let me try to emphasize the point that I was making, because I am not suggesting that one go out of here saying that you never have preset goals or that you are always updating baselines. The distinction that we built on from this morning was distinguishing simple, complicated, and complex situations, where what is simple is where we know how to produce an outcome and we agree as a global community that that outcome is important, like the eradication of polio. It is perfectly appropriate, indeed it would be inappropriate, invalid, and unethical, not to have a clear specific smart goal that polio ought to be eradicated, and the definition of that is very clear. There is no polio anymore, it is gone, no kids are getting polio. That is as clear and specific an outcome as you can get. It is attainable. The world is spending more than a million dollars per case now to make sure that it is attainable, and we are very close, but that means we know how to produce that outcome because we have a technology that will do it, and the world has agreed that that is something we ought to do. That defines the conditions under which you have clear, specific, and measurable outcomes and hold people accountable for them. The World Health Organization has a campaign predicated on a theory of change that will attain that outcome.

Poverty reduction is not polio. There is no vaccine for it. We do not know how to bring about poverty reduction. It takes many different forms. It manifests itself in many different contexts. It is constantly a changing phenomenon affected by very different kinds of contexts. So what I am suggesting, based upon complexity science, is that to act like we are in the simple situation that we know how to produce an outcome and set predetermined goals, when we do not know how to attain those, is fantasy life. It is doing what the complexity theorists are saying one ought not do in the face of complexity, and that is think that you can control it, use mechanisms of control. Setting predetermined goals and having rigid baselines is a command and control strategy. It's not an adaptive strategy. So I am not saying you always update baselines, but that you do so in complex dynamic situations, and you have all experienced this in programs that you run.

One of the common findings that I see in ex-post reports is when people look at why they changed, what they were changing—and I have been a program director at of a big USAID program and experienced this myself—is that when they look at the changes they made, what I hear all the time is, "We actually didn't really understand the problem when we started." That to me is saying we got the baseline wrong. It is not just a matter of understanding. We did not know what the right questions were. Not only did we not have good data, we were not even asking about the right data, we did not understand the situation.

Situation analysis is often done in a fairly abstract way. It is done fairly removed from the situation, and the evidence of complex, unfolding dynamics is that when you get on the ground and start doing stuff, you actually find out what that baseline situation was. Now I take Ali's point about it is manipulable and corruptible, and that is why one needs independent evidence about whether or not those updates are valid and appropriate, but the alternative is to continue to live in the fantasy world that those made up baselines had any meaning. So, we are between a rock and a hard place, we hold on to what we know are not real baselines or we update them and take the risk that that is done badly and without accountability. Find the sweet spot in the middle, which is doing valid and rigorous updating so that we understand what the situation was to some extent.

That is what I'm talking about, as well as emergent goals. The key thing here, the overall message that I hope folks are taking away in conjunction with my colleagues throughout the morning, is situational appropriateness for evaluation itself: that we do different kinds of evaluation for different situations, different knowledge areas, different degrees of change, different kinds of problems. And a recognition that evaluation as currently practiced has been for one kind of situation, one kind of understanding about how the world has changed, and that complexity and crisis present different kinds of situations that present different challenges for evaluation, and that imposing our traditional practices on those new situations is going to not only not work very well, but actually can do damage, can do harm, because it stops programs from adapting in ways that they need to do. We impose rigidities, we impose mechanistic models on programs that need to be adaptive and need to be changing.

So this is not just an abstract kind of thing, it is not that stakes are not very high. What has gotten me impassioned about this is seeing evaluation do so much damage by keeping programs from being able to adapt and do a better job because of narrow, simpleminded kind of accountability constraints. So that is the message here: What is the real situation? How do we define that? That is the interpretive mindset. How do you define the situations you are getting into, and then do you have evaluation approaches that can be adaptive to those different situations? One size does not fit all.

Patrick G. Grasso

Michael, thank you very much. Michael once recited a little aphorism, which is that you do not need a randomized control trial to demonstrate that jumping out of a plane without a parachute is a very bad idea. I would say you do not need a randomized control trial to evaluate this day's session as a real success. I think the people who organized it ought to be congratulated, and our speakers ought to be thanked. So, thank you.

ANNEX

UTILIZATION-FOCUSED EVALUATION: REAL-TIME AND PROSPECTIVE ASPECTS

Michael Quinn Patton

Utilization-focused evaluation is evaluation done for and with specific primary intended users for specific, intended uses. Utilization-focused evaluation begins with the premise that evaluations should be judged by their utility and actual use; therefore, evaluators should facilitate the evaluation process and design any evaluation with careful consideration for how everything that is done, from beginning to end, will affect use. Use concerns how real people in the real world apply evaluation findings and experience the evaluation process. Therefore, the focus in utilization-focused evaluation is on achieving *intended use by intended users*. In responding to the challenges of the real-time and prospective aspects of evaluation, utilization-focused evaluation includes an option I call *developmental evaluation*, where the intended use is development under conditions of complexity. I shall argue that this is a distinct and important evaluation purpose. The primary intended users are social innovators and others working to bring about major systems change (Patton, 2008).

AN OVERVIEW OF UTILIZATION-FOCUSED EVALUATION

In any evaluation there are many potential stakeholders and an array of possible uses. Utilization-focused evaluation requires moving from the general and abstract, i.e., possible audiences and potential uses, to the real and specific: actual primary intended users and their explicit commitments to concrete, specific uses. The evaluator facilitates judgment, decision making, and action by intended users. Developmental evaluation, conducted from a utilization-focused perspective, facilitates ongoing innovation by helping those engaged in innovation examine the effects of their actions, shape and formulate hypotheses about what will result from their actions, and test their hypotheses about how to foment change in the face of uncertainty in situations characterized by complexity.

Utilization-focused evaluation is personal and situational. The evaluation facilitator develops a working relationship with intended users to help them determine what kind of evaluation they need. This requires negotiation in which the evaluator offers a menu of possibilities within the framework of established evaluation standards and principles. Thus, while concern about utility drives a utilization-focused evaluation, the evaluator must also attend to the evaluation's accuracy, feasibility, and propriety (Joint Committee on Standards, 1994). Moreover, as a professional, the evaluator has a responsibility to act in accordance with the profession's adopted principles of conducting systematic, data-based inquiries; performing competently; ensuring the honesty and integrity of the entire evaluation process; respecting the people involved in and affected by the evaluation; and being sensitive to the diversity of interests and values that may be related to the general and public welfare (AEA, 2004).

Utilization-focused evaluation does not advocate any particular evaluation content, model, method, theory or even use. Rather, it is a process for helping primary intended users select the most appropriate content, model, methods, theory and uses for their particular situation. Situational responsiveness guides the interactive process between evaluator and primary intended users. Developmental evaluation is one of the options now available in the feast that has become the field of evaluation. Utilization-focused evaluation can include any evaluative purpose (formative, summative, developmental), any kind of data (quantitative, qualitative, mixed), any kind of design (e.g., naturalistic, experimental) and any kind of focus (processes, outcomes, impacts, costs, and cost-benefit, among many possibilities). Utilization-focused evaluation is a process for making decisions about these issues in collaboration with an identified group of primary users focusing on their intended uses of evaluation.

A psychology of use undergirds and informs utilization-focused evaluation. In essence, research and my own experience indicate that intended users are more likely to use evaluations if they understand and feel ownership of the evaluation process and findings; they are more likely to understand and feel ownership if they've been actively involved; and by actively involving primary intended users, the evaluator is training users in use, preparing the groundwork for use, and reinforcing the intended utility of the evaluation every step along the way. Developmental evaluation carries this user involvement farther than usual by creating a dynamic partnership between social innovators and the developmental evaluator. The language of "partnership" is not the norm in describing the relationship between an evaluator and those whose work is being evaluated. Thus, developmental evaluation invites both skepticism and controversy.

SITUATION RECOGNITION

Astute situation recognition is at the heart of utilization-focused evaluation. There is no one best way to conduct an evaluation. This insight is critical. The design of a particular evaluation depends on the people involved and their situation. The standards and principles of evaluation provide overall direction, a foundation of ethical guidance, and a commitment to professional competence and integrity, but there are no absolute rules an evaluator can follow to know exactly what to do with specific users in a particular situation. Recognizing this challenge, situation analysis is one of the essential competencies for program evaluators.

The idea – admittedly an ideal -- is to match the type of evaluation to the situation and needs of the intended users to achieve their intended uses. This means – and I want to emphasize this point – *developmental evaluation is not appropriate for every* *situation*. Not even close. Indeed, I shall argue that its niche is small and demanding. It will not work if the conditions and relationships are not right. I'll be specifying what those conditions and relationships are as we proceed. The point here is that every evaluation involves the challenge of matching the evaluation process and approach to the circumstances, resources, timelines, data demands, politics, intended users, and purposes of a particular situation. Matching requires astute situation recognition.

Distinguishing simple, complicated, and complex situations

To facilitate situation recognition, it is useful to have a heuristic framework, some way of cutting to the chase by knowing what factors are important to consider when we encounter a new situation. Heuristics are short-cuts that tell us what's important to pay attention to. We cannot look at everything. We never have perfect information. We can't consider all possibilities. We need some way of focusing. Heuristics do that. Research on decision-making shows that heuristics "make us smart" – smart in the sense that we make intelligent decisions quickly. Heuristics direct us in making sense of things. They frame and inform decisions. Indeed, they make choices and action possible.

Developmental evaluation is especially appropriate for complex situations and aims to inform fast action and quick reactions by social innovators. First, then, we have to decide if we're in a situation that is appropriate for developmental evaluation, a complex situation, where the pace of actions, reactions, and interactions matter greatly. In writing the book *Getting to Maybe: How the World Is Changed* (Westley, Zimmerman, & Patton 2006) we looked at the implications of these distinctions for understanding social innovation. In this paper I want to apply these distinctions to illuminate evaluation situations and options.

Remember, the focus here is on utility. These distinctions help with situation recognition so that an evaluation approach can be selected that is appropriate to a particular situation and intervention, thereby increasing the likely utility -- and actual use – of the evaluation. Using these distinctions involves mapping the territory and context within which an evaluation will take place to locate the evaluation within that territory. Moreover, these are relative and perspective-dependent distinctions, not absolute. A situation can be described as more or less simple, complicated, or complex. Utility resides in examining the implications and insights generated by asking to what extent a situation is usefully approached as simple, complicated, or complex, or some combination.

The Degree of Uncertainty/Degree of Conflict Matrix

The degree of uncertainty/degree of conflict matrix developed by Zimmerman (adapted from ideas of Ralph Stacey as published in Zimmerman, Lindberg, & Plsek, 1998, pp. 136-143) is the basis for the heuristic used here that distinguishes simple, complicated, and complex situations. To make these distinctions, the matrix maps the situation along two dimensions. One dimension scales the degree of certainty about what should be done to solve a problem. We know how to eradicate polio. Immunize all children. We don't know how to reduce global warming. There are many competing ideas and plans, but, in fact, our knowledge is quite limited about both the causes of global warming and what interventions would work. Programs and interventions are close to certainty when the cause and effect relationship is highly predictable, as in the relationship between vaccination and preventing disease. At the other end of the certainty continuum are innovative programs where the outcomes are highly unpredictable. Comprehensive anti-poverty initiatives involve considerable uncertainty. Extrapolating from past experience is problematic because each community is unique and there is no immunization against poverty.

FIRST HEURISTIC DIMENSION

Degree of *certainty and predictability* about how to solve a problem

Close to certainty \\---/ Far from certainty

The second dimension depicts the degree of agreement among various stakeholders about an intervention's desirability, or alternatively, the degree of conflict. There is universal agreement that preventing polio is a good thing and that children should be vaccinated to eradicate polio worldwide. On the other hand, there is substantial political conflict about almost all aspects of global warming. To what extent is global warming occurring? To what extent is it caused by human activity (as opposed to being a natural earthly cycle)? What are the primary causes of climate change? How much urgency is there about intervening? What interventions, if any, will make a difference? Are the economic costs of intervening worth the likely results? On these and other matters, there is great disagreement.

SECOND HEURISTIC DIMENSION

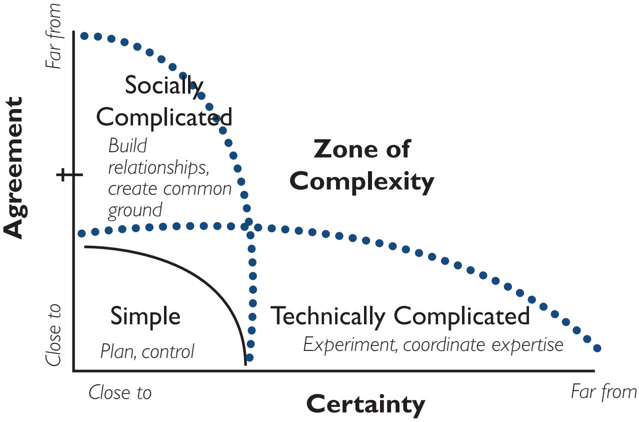
Degree of *agreement* or *conflict* about how to solve a problem

Close to agreeing \\---/ Far from agreeing
Little conflict Great conflict

Combining these two dimensions creates the borders of a territory that can be mapped, or a matrix, as shown in Exhibit 1. The horizontal axis captures the degree of certainty and predictability about how to solve a problem. The vertical axis displays the degree of agreement about what to do.

EXHIBIT 1

Know When Your Challenges Are In the Zone of Complexity



Simple situations

High levels of certainty and agreement make situations fairly simple. *Simple*, as used here, is *a descriptive term*, not meant to be judgmental or pejorative. Simple is not simplistic or simple-minded. A simple situation is, simply, one in which knowledge and experience tell you what to do and there is great agreement about what to do. In such a situation, it is both possible and appropriate to intervene from the top-down, as in the worldwide campaign to eradicate polio. The high degree of predictability and agreement permits detailed planning, controlled execution, and precise measurement of the degree to which predetermined targets are reached. A best practice model can be generated and subjected to a summative test.

A *simple* problem is how to bake a cake, a metaphor for the capturing the characteristics of the simple originally offered by Zimmerman and Glouberman (2004). A good recipe, like a best practice, provides detailed guidance about the steps to follow to achieve a desired outcome. A recipe has clear cause and effect relationships and can be mastered through repetition and honing basic skills. Recipes present standard procedures and should provide sufficient detail that even someone who has never baked has a high probability of success. In simple situations, what needs to be done is *known*. Best practices for programs are like recipes in that they provide clear and high fidelity directions. The standard procedures that have worked to produce desired outcomes in the past are highly likely to work again in the future. Assembly lines in factories have a "recipe" quality as do standardized school curricula. Part of the attraction of the 12-Step program of Alcoholics Anonymous is its simple formulation (which doesn't mean it is easy to do, even one day at a time).

Complicated situations

As situations become less predictable and producing desired outcomes becomes less certain, we are moving into *complicated* territory. It is useful to distinguish technical complications from social complications. Sending a rocket to the moon is technically *complicated* because there are thousands of elements that have to be coordinated for a successful launch. Technical knowledge and expertise is needed to solve complicated problems. More than one area of expertise is needed and must, therefore, be coordinated and integrated. In rocket science, formulae are used to predict the trajectory and path of the rocket. Calculations are required to ensure sufficient fuel based on current conditions. If all of the many technical calculations are done well, coordinated, and executed precisely, it is likely that the desired outcome – getting the rocket to the moon – will be accomplished. Like integrating the many areas of expertise needed to get a rocket into space, coordinating large-scale programs with many local sites throughout a country or region is a complicated problem. When the degree of uncertainty and agreement are such that what needs to be done is challenging and difficult, but *knowable*, the situation is complicated. That is, how all the parts will fit together is initially unknown but can be figured out, and is therefore knowable, in complicated situations.

Socially complicated situations involve situations with many different stakeholders offering different perspectives, articulating competing values, and posing conflicting solutions. Whether resources should be spent sending rockets into space is more controversial than whether polio should be eradicated worldwide, thus rocket launches are more socially complicated than immunization campaigns (at least for purposes of illustrating the conceptual difference between simple and complicated). Abortion is an example of a socially complicated issue, as is what to do about the energy crisis. Everyone wants children to learn to read but there are intense disagreements about which reading approach produces the best result. Controversial issues like sex education are socially complicated. The more points of view there are and the greater the debate among different stakeholders, the more socially complicated the situation becomes. How diverse stakeholders will deal with their conflicts is initially unknown but knowable as the interactions unfold. Some of the disagreements may be about degree of technical complication (how much certainty there is about how to produce a desired outcome), but many disagreements are about fundamental value differences and how to even define the problem.

Having distinguished the technically complicated from the socially complicated and given illustrations of each, we need to combine them to look at their interactions. A situation is **complicated** when there is either a high degree of uncertainty <u>or</u> a high degree of disagreement. If there is *both* high uncertainty and high disagreement (for instance, uncertainty is a primary source of disagreements and disagreements contribute to the uncertainty), we have moved into the arena of complexity.

Complex situations

Complex situations are characterized by high uncertainty and high social conflict. In studying social innovations, we were impressed by the uncertainty and unpredictability of the innovative process, even looking back from a mountaintop of success, which is why we called the book *Getting to Maybe* (Westley, Zimmerman and Patton 2006). Evaluating social innovations is a complex challenge, as opposed to evaluating simple and complicated problems. The outcomes of interventions aimed at solving problems under conditions of complexity are unpredictable. So many factors and variables are interacting, many of them not only unknown but *unknowable*, that there can be no recipe for success. And even if something that looks like a recipe emerges from one or two successful attempts to do something, the likelihood that the same result can be attained in other and different contexts is low. There are simply too many dynamic variables and unknowns to make recipe-like replication (or supposed best practices) predictable.

It's worth reiterating the interactions between high uncertainty and high disagreement. These interaction are volatile, uncontrollable, unpredictable, and unknowable in advance: *high uncertainty about how to produce a desired result fuels disagreement, and disagreements intensify and expand the parameters of uncertainty.*

Parenting is *complex*. Unlike the simple metaphor of a cooking recipe or the rocket launching metaphor for a complicated situation, parenting involves huge uncertainties and no clear rules guaranteeing success to follow. Oh, to be sure, there are many experts in parenting and many guides available to parents. But none can be treated like a cook book for a cake, or a set of formulae to send a rocket to the moon. In the case of the cake and the rocket, for the most part, we were intervening with inanimate objects. The flour does not suddenly decide to change its mind and gravity can be counted on to be consistent too. On the other hand, children, as we all know, have minds of their own. Hence our interventions are always in relationship with them. There are very few stand-alone parenting tasks. Almost always, the parents and child interact to create outcomes.

CAUSE AND EFFECT RELATIONSHIPS

At the heart of the distinctions between simple, complicated, and complex is the extent to which cause and effect is or can be known. In simple situations cause and effect is known so interventions and their consequences are highly predictable and controllable. In complicated situations cause and effect is knowable as patterns are established through research and observations over time, but the many variables involved make prediction and control more precarious. In complex situations, cause and effect is unknown *and* unknowable until after the effect has emerged, at which point some retrospective tracing and patterning may be possible. These different degrees of *causal knowability* actually define the uncertainty dimension of the degree of uncertainty/degree of conflict matrix. Causal knowability is a distinguishing element distinguishing simple, complicated, and complex. Management and organizational development consultant David Snowden has emphasized these different degrees of causal clarity to distinguish simple, complicated, and complex, with special attention to their implications for management planning and action (Snowden and Boone, 2007).

The *Cynefin* Framework

> *Wise executives tailor their approach to fit the complexity of the circumstances they face.*
>
> Snowden and Boone (2007, p. 68)

This was the central message of "A Leader's Framework for Decision Making" by management consultants David Snowden and Mary Boone in their featured *Harvard Business Review* article. The article was designated as the Best Practitioner-Oriented Paper in Organizational Behavior in 2007 by the Organizational Behavior Division of the Academy of Management. As Brenda Zimmerman was refining the distinctions between simple, complicated and complex in the certainty and agreement matrix, David Snowden and colleagues in IBM's Institute of Knowledge Management were thinking in parallel terms that led to the *Cynefin* framework, making the same distinctions, an impressive exemplar of independent discoveries by creative minds following the same path.

Snowden, of Welsh lineage, chose the Welsh word **Cynefin** (pronounced kun-ev'in) as the name of the framework distinguishing simple, complicated, complex, and chaotic. The Welsh dictionary translates *cynefin* as meaning haunt, habitat, acquainted, accustomed, or familiar, being both noun and adjective, and thus requiring context to understand its meaning in any given instance. Snowden resonated to this uncertainty which evokes the sense that our understandings depend on our interactions with each other and our environment, which includes cultural traditions, organizational norms, and the geographical/ecological setting within which interactions occur. Snowden's *cynefin* framework emphasizes variations in the nature of causality and the corresponding implications for decision-making and action (Snowden and Boone ,2007; Kurtz and Snowden, 2003).

Simple: linear, direct connection between cause and effect; easily observable, understandable, and verifiable. This is the arena where things are *known*, so best practices can be identified and applied. A leader's or manager's decision/action sequence is:

Sense → Categorize → Respond

Complicated: determining cause and effect requires analysis and expert investigation, so things are not yet known, but are knowable. Good, effective practices can be identified (but *not* "best"). The decision/action sequence is:

Sense → Analyze → Respond

Complex: Cause and effect is contingent on contextual and dynamic conditions, and therefore unknowable; patterns are unpredictable in advance. Practice is emergent and contingent. A leader's or manager's decision/action sequence should be:

Probe → Sense → Respond

Chaotic: no observable or predictable relationship between cause and effect because of rapidly changing and highly unstable/turbulent systems dynamics, but some kind of action is required. The appropriate decision/action sequence is:

Act → Sense → Respond

New Zealand evaluator and leading systems thinker Bob Williams (cf. Williams & Iman, 2007) shared with me his experience using the *cynefin* framework. I was exploring a new method of handling patients within a healthcare situation. I got people to group those aspects of the situation into Snowden's four categories (simple/known, complicated/knowable, complex/unknowable, chaotic), acknowledging that a given situation has elements of all four states (each of which implies a different response - including strategies that might move an aspect of the situation form one "state" to another and thus make it easier to manage).

This then leads to some very interesting conversations about whether they were assuming that a problem was "knowable" if only they worked hard enough, or that they were looking for "best practice" when actually "good practice" was what they should be considering. Some aspects of the situation were placed in more than one category. At this point all kinds of light bulbs lit up. People realized that part of the problem they were experiencing was that different people were imagining that aspect from two different understandings of what is going on. They suddenly understood why they were having difficulty resolving or managing the situation: "Oh so you were managing it as if it were complicated and I was managing it as if were complex - no wonder we were clashing over strategies."

Snowden's focus has been on teaching leaders and managers to make *cynefin* framework distinctions as a guide to decision-making. My focus here is on its implications for evaluators. Exhibit 2 adapts his Leader's Guide to Decisions in Multiple Contexts to evaluation.

EXHIBIT 2
Decisions in Multiple Contexts: An Evaluator's Guide

Wise evaluators tailor their approach to fit the complexity of the circumstances they face.

	The Situation: Agreement/Certainty Matrix and *Cynefin* Framework	The Leader's Job
SIMPLE	High agreement about the problem and what to do; high certainty that the right action will produce the desired results: clear, direct, linear, predictable, and controllable cause-effect pattern. What needs to be done is known.	Sense, categorize, respond. Know what is known. Manage based on facts. Advocate for and implement best practices.
COMPLICATED	Some disagreements about the problem and what to do. Expertise needed. The necessity of coordinating many areas of technical expertise and many actors introduces uncertainty about attaining desired outcomes. More than one effective way possible. Cause-effect linkages are context-contingent; discoverable with careful analysis, but neither obvious nor certain. Contingencies discernible (known unknowns).	Sense, analyze, respond. Find needed expertise to identify good practices. Listen to and assess conflicting expert advice. Use monitoring and evaluation to track what unfolds as good practices are tried.
COMPLEX	High uncertainty about how to produce desired results and great disagreement among diverse stakeholders about the nature of the problem and what, if anything, to do. Results highly dependent on initial conditions; non-linear interactions within a dynamic system. No right answers; key variables and their interactions unknown in advance. Each situation is unique.	Probe, sense, respond. Foster dialog, creativity and innovation. Watch for and interpret emerging patterns. Be flexible and adaptive. Make time for and engage in reflective practice to capture, understand, and interpret what is emerging.
CHAOTIC	High conflict among stakeholders; extreme uncertainty about what to do. Turbulence and volatility make pattern detection unreliable, even undecipherable. Dynamic interactions hard to follow, not even sure what to pay attention to. Unreliable information. What to focus on is unknown and a matter of great debate. Tense, stressful decision environment.	Act, sense, respond. Try things out and see what happends, watching for anything that works. Manage what is manageable to establish some degree of order. Don't yield to panic.

Source: Patton (2010).

The Evaluator's Job	Evaluation Challenges
Validate best practices (summative evaluation). Monitor implementation of best practices to assure high fidelity, adherence, and quality. Report departures from best practices amd implications of those departures, especially implications for outcomes.	Assuring that best practices fit new contexts (different from where the practices were originated and validated). Detecting unanticipated consequences and context-specific implementation problems.
Validate effective practices and options with attention to context and system contingencies. Convert expert advice into a testable theory of change. Evaluate and report unfolding cause-effect complications and their implications. Systems thinking.	Designing a reasonable test of the theory of change (summative evaluation). Understanding the system(s) and context(s) within which action unfolds. Detecting and measuring both outcomes and contingencies. Facilitating interpretation of less-than-certain findings.
Identift and document initial conditions and monitor what emerges. Provide ongoing, timely, and rapid feedback about what is emerging. Track incremental actions and decisions that affect the paths taken (and not taken). Facilitate regular reflective practice about what is *developing.* Embed evaluative thinking in the innovative process.	Keeping up with the rapid pace of change in turbulent and dynamic environments, and documenting developments. Managing a flexible, emergent design. High level of ongoing interaction and communication. Combining creative and critical (evaluative) thinking in support of innovation. Facilitating interpretation of emergent findings for action. Staying developmentally focused.
Distinguish better and worse data; some information may be better than none, but interpret cautiously. Find those parts of the action where evaluation can make an immediate contribution to help survive chaos.	Acknowledging data inadequacies. Being open and opportunistic about finding data. Avoiding defaulting to the simple in an effort to exercise control and create the illusion of certainty where none exists. Helping to transition to stability in the face of chaos. Don't be a burden.

APPLYING COMPLEXITY CONCEPTS TO REAL-TIME AND PROSPECTIVE ASPECTS OF EVALUATION

The basic premise here is that evaluation in complex adaptive systems is more likely to be useful if the evaluation is informed by complexity concepts and understandings. Pretty straightforward premise -- derived from the importance of matching the evaluation to the nature of the situation. While complexity ideas raise doubts about linear, formulaic, and mechanical models of the world, controversies surround complexity constructs, raising doubts about whether agreement can ever be reached on core constructs. What is not in doubt is that complexity ideas are in vogue, have a lot of currency these days, and, thereby, have attracted ardent adherents and fervent critics.

What brings me to complexity is its utility. It identifies a set of intervention circumstances that are amenable to a particular situationally appropriate evaluation response, what I am calling here developmental evaluation. *Complexity is a defining characteristic of developmental evaluation's niche*. Principles for operating in complex adaptive systems inform the practice of developmental evaluation. The controversies and challenges that come with complexity ideas will also and inevitably afflict developmental evaluation. The insights and understandings of complexity thinking that have garnered enthusiasm from social innovators will also envelope developmental evaluation and open pathways for increasing the credibility, relevance, and utility of evaluation undertaken from a specifically developmental perspective.

Ramalingam and Jones (2008), in a comprehensive review of the application of complexity theory to international humanitarian aid, distinguish three points of view about complexity theory: champions, critics, and pragmatists. Their description of pragmatists nicely summarizes my own perspective, so I cite it here:

> The pragmatists, for whom complexity provides interesting and potentially useful parallels, are exploring the relevance of complexity science to social systems and organisations, and working to assess the practical benefits that arise from its application outside the natural sciences…. This work suggests that complexity is a lens that helps us look at our world and shape our action but, importantly, that it is a set of concepts and tools that should not be treated as the 'only way' to look at and do things. The pragmatists tend to accept the work-in-progress nature of complexity sciences, and the challenges that arise from drawing on diverse and varied bodies of knowledge. These challenges create issues around definition, measurement, analysis and coherence, and lead to a general acknowledgement that there is a need for a deeper theoretical understanding and further practical applications. (Ramalingam & Jones, 2008, p.6)

So, from a pragmatic perspective, what are some of the compelling complexity constructs that inform developmental evaluation? I've focused on six central complexity ideas: nonlinearity, emergence, adaptation, uncertainty, dynamical systems change, and co-evolution. Exhibit 3 defines each of these concepts and suggests their implications for developmental evaluation.

FIVE DEVELOPMENTAL EVALUATION PURPOSES AND USES

In considering the relevance of systems thinking and complexity concepts for evaluation, I want to suggest that developmental evaluation is particularly appropriate for but needs to be matched to five different complex situations and developmental purposes:

1. *Ongoing development* in adapting a project, program, strategy, policy, or other innovative initiative to new conditions in complex dynamic systems.
2. *Adapting effective general principles to a new context* as ideas and innovations are taken from elsewhere and developed within a new setting, the work of developmental evaluation in the dynamic middle between top-down and bottom-up forces of change.
3. *Developing a rapid response* in the face of a sudden major change or a crisis, like a natural disaster or financial melt-down, exploring real-time solutions and generating innovative and helpful interventions for those in need.
4. *Pre-formative development of a potentially scalable innovation* to the point where it is ready for traditional formative and summative evaluation; pre-formative developmental evaluation works with emerging ideas and visionary hopes in a period of exploration to shape them into a potential model that is a more fully conceptualized, potentially scalable intervention. (As models emerge out of exploratory and innovative initiatives, some may move into more traditional formative and summative evaluation to determine scalability and generalizability, while others remain in developmental mode, either undergoing further development or continuous experimentation in the search for new models.)
5. *Major systems change and cross-scale developmental evaluation,* providing feedback about how major systems change is unfolding, evidence of emergent tipping points, and/or how an innovation is or may need to be changed and adapted as it is taken to scale, that is, as its principles are shared and disseminated in an effort to have broader impact. Horizontal scaling across systems or vertical scaling to broader systems may involve more than adaptation; these dissemination and scaling processes can evolve an essentially new development, the emergence of which can be documented and analyzed as part of a developmental evaluation.

ISSUES IN REAL-TIME AND PROSPECTIVE ASPECTS OF UTILIZATION-FOCUSED EVALUATION

Real-Time versus Developmental Evaluation

Real time refers generally to rapid feedback and response, linking data and action as close together in time as possible. The ultimate in real-time data analysis is reporting on stock market transitions in micro-seconds. In hospitals, real time means getting blood analyses or other diagnostic tests back to a doctor within a short timeline that can range from minutes to an hour. In evaluation situations, real time typically means getting

EXHIBIT 3
Characteristics of Complex Systems and Implications for Developmental Evaluation

Characteristics of Complex Systems	Implications for Developmental Evaluation
1. **Nonlinear:** Sensitivity to initial conditions; small actions can stimulate large reactions, thus the *butterfly wings metaphor* (Gleick, 1987); *black swans* (Taleb, 2007), in which highly improbable, unpredictable, and unexpected events have huge impacts; and tipping points (Gladwell, 2002) when major shifts occur changing the whole landscape of action.	Watch for, sample, and study critical incidences. Assess and map tipping points and other changes in the intervention landscape. Use mixed methods to capture when cumulative quantitative changes in key indicators become substantively significant qualitative shifts. Don't confuse linear logic models and strategic plans with what actually goes on in programs. Look for contextual changes that shift program patterns, forks in the road that move the program in new directions, and sudden (or gradual) responses to unexpected developments.
2. **Emergence.** Patterns emerge from self-organization among interacting agents. Each agent or element pursues its own path but as that path intersects with, and the agent interacts with others, also pursuing their own paths, patterns of interaction emerge and the whole of the interactions cohere, becoming greater than the separate parts. What emerges can be beyond, outside of, and oblivious to any notion of shared intentionality (Johnson, 2001).	Be especially alert to formation of self-organizing subgroups who have different experiences of the program and, correspondingly, different outcomes. Anticipate and expect emergent issues and take seriously the search for unanticipated consequences, tracking interactions among key players, both formal and informal, planned and unplanned. Map networks, system relationships, and subgroups. Track information flows, communications, and emergent issues. Emergence applies to both processes and outcomes. Watch for and assess not only what emerges, but what declines or even disappears. Disappearance is the other side of the phenomenon of emergence. The unplanned emerges; the planned disappears. Both are important, as is what unfolds as planned. The evaluation design is also emergent.
3. **Adaptive:** Interacting elements and agents respond and adapt to each other, and to their environment, so that what emerges is a function of ongoing adaptation both among interacting elements and the responsive relationships interacting agents have with their environment. Adaptive management is a systematic, iterative process for making decisions in the face of uncertainty, reduced control, and low predictability, through ongoing system monitoring and response to changes in context. The process essentially involves learning by doing and observing, then making adjustments based on what has been learned, and repeating this cycle of sensing, learning, and adapting over and over.	Regularly capture perspectives from key actors in different but interacting systems about what's going on. Put these perspectives in dialogue with each other to capture and track adaptations and their significance. Both new processes and new outcomes may emerge requiring new evaluation design elements and measures. *The evaluation itself must be adaptive.* An adaptive mindset essentially involves learning by doing and observing. This parallels the process recommended by knowledge management consultant David Snowden when facing complexity: probe, sense, respond Snowden & Boone, 2007). Probing is the doing. Sensing is the observing (where chance ever favors the prepared mind). And responding is the adaptation. The feedback provided by the developmental evaluator informs the innovators' adaptive process, including heightening awareness of what incremental adaptations are occurring so that learnings can be identified and captured. The evaluator may also point out when innovators are not being adaptive despite what is emerging; or when there is increasing uncertainty within a system but the innovators are behaving as if they've figured things out and know what is happening.

4. **Uncertainty.** Under conditions of complexity, processes and outcomes are unpredictable, uncontrollable, and unknowable in advance. Emergent and adaptive self-organization can create idiosyncratic bumps in patterns that becomes mounds that sometimes go on to become idiosyncratic mountains, or at other times erode into nothingness, and it's impossible to know ahead of time which pattern, if either, will prevail. Not acknowledging and dealing with uncertainty and unexpected events can lead to a spiral of disruption with things getting worse (Weick & Sutcliffe (2001, p. 2). Uncertainty is a defining characteristic of complexity. (Westley, Zimmerman, & Patton, 2006).	Identify and acknowledge sources of uncertainty, including: inadequate knowledge about how to produce desired outcomes; disagreements among key actors about what to do, including value conflicts; and turbulence in the larger environment. Work with key stakeholders and primary intended users on an ongoing basis to understand the implications of uncertainty. Nurture tolerance for ambiguity and messiness. This means resisting the temptation to address uncertainty by imposing order and control through evaluation by forcing the complex into a simple linear evaluation logic model with predetermined clear, specific, and measureable outcomes. Provide rapid feedback about unexpected events and their implications. Early detection of and feedback about emergent patterns can be critical. In early stages of trouble or opportunity, the unexpected may give off *weak signals.* "The overwhelming tendency is to respond to weak signals with a weak response." Understanding the potential significance of weak signals and responding strongly "holds the key to managing the unexpected" (Weick & Sutcliffe, 2001, p. 4).
5. **Dynamical:** Interactions within, between, and among subsystems and parts within systems can be volatile, changing rapidly and unpredictably due to the interdependence of key factors and variables. The system may shift from rest to rhythmic oscillation to random thrashing. These changes seem to be spontaneous, but they are driven by the internal dynamics of the system itself as the constraining conditions interact with each other to influence the behaviors of agents in the system.	Track and document not only whether change occurs, but how and why it occurs. Processes and outcomes can be both dynamic and dynamical; pay attention to both, and their interrelationship. Create a flexible and responsive data collection system that can mirror adaptive, emergent, and dynamic/dynamical developments, so that fieldwork can speed up and slow down in sync with the intervention's rhythms of change. Engage in ongoing monitoring of shifts in levels of activity to capture dynamic/dynamical transitions. Analyze and distinguish contextual factors and participation patterns that are static, dynamic, and dynamical, and the implications of these different patterns.
6. **Co-evolutionary:** As interacting and adaptive agents self-organize, ongoing connections emerge that become *co-evolutionary* as the agents evolve together (co-evolve) within and as part of the whole system, over time.	Developmental evaluation will co-evolve with the innovation and intervention, both affecting innovation and being affected by it. This is a process of co-creation. The evaluation will not be independent and separate from the innovation but will be interdependent with it, and with those involved in it (as part of a team), as the evaluator provides feedback, facilitates conceptualization of the change process, and both captures and generates perspectives about what is happening, and why. *Process use,* in which evaluative thinking affects the intervention, will be as important as findings use.

results to intended users in a day or two, or at most a couple of weeks, rather than in months or on a routine schedule of standard quarterly reports (a common information system reporting timeframe).

Developmental evaluation aims for real-time feedback, but *not all real-time data use and evaluation is developmental.* Police departments use real-time data on increasing crime in a neighborhood to reallocate personnel from lower crime to higher crime areas. That is real-time evaluation and data use, but it is not developmental. This real-time use of data by police involves implementing a rapid response management approach, but the police are not developing that approach. In contrast, if crime data in a community indicated a national gang was moving into the community, the police could develop a task force to fight gang recruitment, infiltration, and crime and monitor emergent effects as the gang adapted to police attention so that police could adapt accordingly. That would be developmental evaluation because the intervention is emerging in real time and using evaluation data to adapt the intervention to what emerges in real time.

Developmental Evaluation versus Development Evaluation

Developmental evaluation is easily confused with development evaluation. They are not the same, though developmental evaluation can be used in development evaluations. This has created some confusion, which I regret, and hereby address.

Development evaluation is a generic term for evaluations conducted in developing countries, usually focused on the effectiveness of international aid programs and agencies. The work of IEG is development evaluation. *The Road to Results: Designing and Conducting Development Evaluations* (Imas & Rist, 2009) is an exemplar of this genre, a book based on The World Bank's highly successful International Program for Development Evaluation Training (IPDET) which the book's authors founded and direct, and on which their book is based. Full disclosure: I have been on the IPDET faculty since the program began.

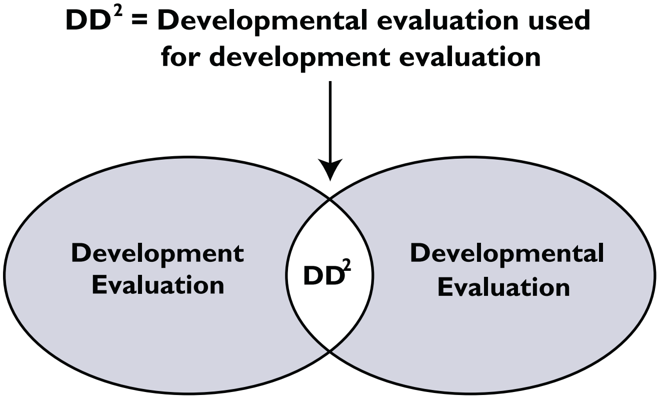
Developmental evaluation, as defined and described in the *Encyclopedia of Evaluation* (Mathison, 2005, p.116), has the purpose of helping develop an innovation, intervention, or program. In developmental evaluation the evaluator typically becomes part of the program or innovation design team, fully participating in decisions and facilitating discussion about how to evaluate whatever happens. All team members, together, interpret evaluation findings, analyze implications and apply results to the next stage of development. The evaluator becomes involved in improving the intervention and uses evaluative approaches to facilitate ongoing program, project, product, staff and/or organizational *development.* The evaluator's primary function in the team is to facilitate and elucidate team discussions by infusing evaluative questions, data and logic, and to support data-based decision-making in the developmental process. In this regard, developmental evaluation is analogous to research and development (R & D) units in which the evaluative perspective is internalized in and integrated into the operating unit. In playing the role of developmental evaluator, the evaluator helps make an intervention's development an R & D activity.

Part of the value of an experienced developmental evaluator to an innovation team is bringing a reservoir of knowledge (based on many years of practice and having read a great many evaluation reports) about what kinds of things tend to work and where to anticipate problems. Experienced evaluators have typically accumulated a great deal of knowledge and wisdom about what works and doesn't work. More generally, as a profession, the field of evaluation has generated a great deal of knowledge about patterns of effectiveness. That knowledge makes evaluators valuable partners in designing as well as evaluating social innovations.

An evaluation focused on development assistance in developing countries could use a developmental evaluation approach, especially if such developmental assistance is viewed as occurring under conditions of complexity with a focus on adaptation to local context. But developmental evaluations are by no means limited to projects in developing countries. Developmental evaluation can be used anywhere that social innovators are engaged in bringing about systems change under conditions of complexity.

The *al* in d*evelopmental* is easily missed, but it is critical in distinguishing development evaluation from developmental evaluation. Exhibit 4 portrays the relationship between development evaluation and developmental evaluation.

EXHIBIT 4


DD² = Developmental evaluation used for development evaluation
Development Evaluation
DD²
Developmental Evaluation

When I first labeled and wrote about *developmental evaluation* 15 years ago (Patton, 1994), development evaluation was not a distinct and visible category of evaluation practice and scholarship. Evaluations in developing countries were certainly being conducted, but an identifiable body of literature focused on evaluating development assistance had not attracted general professional attention. One of the most important trends of the last decade has been the rapid diffusion of evaluation throughout the world, including especially the developing world, highlighted by formation of the International Development Evaluation Association which launched in Beijing, China, in 2002. IEG has been a leader in developing development evaluation as a field of professional practice in evaluation.

Confusion about the distinct and sometimes overlapping niches of development evaluation and developmental evaluation is now, I'm afraid, part of the complex landscape of international evaluation. I hope this paper helps sort out both the distinctions and the areas of overlap.

Ten other issues and controversies in Prospective Evaluation under conditions of complexity

Here are some of the issues and controversies in Prospective Evaluation under conditions of complexity:

1. *Maintaining a results focus*: Should there be and can there be pre-ordinate targeted outcomes? How can interventions be results-oriented under conditions of high uncertain and dynamical complexity?
2. *Comparative analysis*: Can baselines be revised given dynamic and dynamic conditions? Getting beyond static and sacrosanct baselines.
3. *Emergence*: How do we take *emergence* seriously? Getting beyond token attention to "unanticipated consequences."
4. *Flexible designs*: How do we adapt evaluation to complex circumstances with emergent and flexible designs and measures?
5. *Evaluation budgeting*: How do we engage in contingency-based evaluation budgeting?
6. *Poverty focus*: How can evaluate under conditions maintain a focus on poverty when more developed (relatively) countries may have more capacity for rapid adaptability?
7. *Evaluation within a macro systems context:* Climates change and the global economic crisis provide a context within which any particular evaluation will unfold for the foreseeable future. How does evaluation take this larger global context into consideration?
8. *Sustainability Concerns:* Under conditions of complexity sustainability means resilience rather than continuity, yet most traditional approaches to evaluation continue to treat continuity as the criterion for sustainability.
9. *Forward- looking (prospective) uncertainties:* Prospective Evaluation will offer probability estimates under conditions of high uncertainty and little likelihood of being accurate. What form should such estimates take? For example, we will likely know more about factors to worry about than be able to offer actual estimates of results, but results estimates may be expected. Can we use scenario approaches instead of static future estimates? What caveats need to be included in prospective evaluation?

10. *Rapid and ongoing updates*: Traditional evaluation focuses the action on the beginning (baseline), middle (progress report) and end (accountability and summative evaluation). In M & E, monitoring has served program management purposes more than evaluation. How can ongoing evaluation and *updating prospective evaluation scenarios* be built into evaluation under conditions of complexity?

The essence of utilization-focused developmental evaluation

So, bottom line: *How can you tell if an evaluation is truly developmental?* The answer lies in focusing on the evaluation's primary purpose and outcomes: *Is the purpose and focus of the evaluation helping develop something? Is something getting developed? Did something get developed? If so, what? How? With what implications?* The focus of developmental evaluation is on *developing and adapting innovations*.

To borrow an old saying, the proof of the pudding is *in the eating.* Since I distinguish developments from improvements, and position developmental evaluation as different in important ways from formative and summative evaluation, let me offer this cooking metaphor. *Distinguished evaluation theorist and practitioner* Bob Stake has explained: When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative. More generally, anything done to the soup during preparation in the kitchen is improvement-oriented; when the soup is served, summative judgment is rendered by the guests who consume the soup. And what of developmental evaluation in this metaphor?

Developmental evaluation begins when, before cooking, the chef goes to the market to see what vegetables are freshest, what fish has just arrived, and meanders through the market considering possibilities, thinking about who the guests will be, what they were served last time, what the weather is like, and considers how adventurous and innovative to be with the meal. If the chef decides to follow a standard recipe, the situation remains appropriate for formative and summative evaluations based on fidelity to the prescribed recipe. If the chef decides to attempt a new creation, innovate, and develop a new dish especially well-suited for these particular guests in the context of this particular evening, then the situation opens up the possibility for creativity and developmental evaluation. And when a guest and a cook create and concoct a soup together, that co-creation is developmental.

Situational Responsiveness and Developmental Evaluation

This entire paper has been about how we figure out what situation we face so we can engage appropriately. In particular, I have been delineating and refining the niche of developmental evaluation as especially appropriate for interventions and innovations being undertaken under conditions of complexity. Applying David Snowden's advice to leaders, the message of this paper has been:

> *Wise evaluators tailor their approach to fit the complexity of the circumstances they face.*

REFERENCES

American Evaluation Association. (2004). Guiding principles for evaluators. http://www.eval.org/Publications/GuidingPrinciples.asp

Gladwell, M. (2002). *The tipping point: How little things can make a big difference*. Boston: Little, Brown.

Gleick, J. (1987). *Chaos: Making a new science*. New York: Penguin.

Johnson, S. (2001). *Emergence: The connected lives of ants, brains, cities, and software*. New York: Scribner.

Joint Committee on Standards. (1994) *The program evaluation standards*. Thousand Oaks, CA: Sage. http://www.wmich.edu/evalctr/jc/

Kurtz, C. F. & Snowden, D. J. (2003) The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal,* 48 (3), 462–483.

Morra-Imas, L.G. & Rist, R. (2009). *The road to results: Designing and conducting development Evaluations.* Washington, D.C.: The World Bank.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice* 15 (3), 311–20.

Patton, M. Q. (2008). *Utilization-focused evaluation, 4th* ed. Thousand Oaks, CA: Sage.

Patton, M. (2010). *Developmental evaluation: Applying complexity concepts to enhance use and innovation.* New York: Guilford.

Ramalingam, B. & Jones, H. with Reba, T. & Young, J. (2008). *Exploring the science of complexity: Ideas and implications for development and humanitarian efforts.* Working Paper 285. London: Overseas Development Institute.

Snowden, D. J. & Boone, M.E. (2007). A leader's framework for decision making. *Harvard Business Review*. 85 (11), 68–77.

Taleb, N.N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Weick, K. E. & Sutcliffe, K. (2001). *Managing the unexpected: Assuring high performance in an age of complexity.* San Francisco: Jossey-Bass.

Westley, F., B. Zimmerman & M. Q. Patton. (2006). *Getting to maybe: How the world is changed.* Toronto: Random House Canada.

Williams, B. & Iman, I. (Eds.) (2007). *Systems concepts in evaluation*: *An expert anthology*. American Evaluation Association monograph. Point Reyes, CA: EdgePress of Inverness.

Zimmerman, B. & Glouberman, S. (2004). Complicated and complex systems: What would successful reform of Medicare look like? In P.G. Forest, T. McIntosh, & G. Marchildon (Eds.) *Health Care Services and the Process of Change* (pp.21-53). Toronto: University of Toronto Press. Originally published as Discussion paper No. 8. (2002). Ottawa: Commission on the Future of Health Care in Canada

Zimmerman, B., Lindberg, C. & Plsek, P. (1998). *edgeware: insights from complexity ideas for health care leaders*. Irving, Texas: VHA.