Presents

"When Will We Ever Learn: Improving Lives through Impact Evaluation" Policy Recommendations from the CGD Evaluation Gap Working Group

[Transcript prepared from a tape recording]

This event was held in Washington, DC
on May 31, 2006 at 2:00 p.m.

Lawrence MacDonald:Good afternoon. Thank you very much for joining us here today. I'm Lawrence MacDonald, Director of Communication and Policy at the Center for Global Development, and it's really my great pleasure to welcome you to the launch of the report of the CGD Evaluation Gap Working Group, When Will We Ever Learn Improving Lives Through Impact Evaluation? You made a brilliant choice in coming here today because the proposals that you're going to hear about have the potential to transform the development business. It's also the case that this is unlike any launch you have attended in Washington before and that, as you came in, you had an opportunity to pick up a piece of paper that is essentially a call to action. It goes by sort of the eminently Washington wonk kind of name, Call for an International Initiative to Foster Independent Impact Evaluation of Social Sector Programs and Policies. It would take this crowd to find that exciting, but I take some responsibility for that title, so I join you in that. It is not a public petition. We're not asking just anybody to sign it. Rather, it's a call to action by a diverse cross section of the development community. A few of you have signed it already on line. Some of you have signed it. I see there's at least one in the box out in the hallways. Some of you might have decided you were going to sign it already, but I encourage you, if you haven't decided, as you listen to our speakers today, to consider whether or not you would lend your name in support of this initiative. It's now my pleasure to introduce our speakers. We will hear first from my colleague, Ruth Levine. She's the Director of Programs and a Senior Fellow at the Center for Global Development and has been for the last several months, our Acting President. She previously had two jobs and recently, she's had three. In the meantime, in her spare time, she put out this report. Ruth was previously, among other things, responsible for research on the health sector at the World Bank and also for knowledge management activities there, and she served as an advisor on the Social Sectors in the Office of Executive Vice President at the IDB. She is the author of many articles and a number of books, including *The Health of Women in Latin America and the Caribbean, Making Markets for Vaccines – Ideas to Action, and Millions Saved – Proven Successes in Global Health*. After Ruth speaks about the general purpose of this initiative, we'll hear from Bill Savedoff. He's a senior partner at Social Insight. Bill has worked for more than 15 years on improving the quality of social services in developing countries in Africa, Asia and Latin America, and Bill will be speaking about the specific recommendations of the working group. Ruth.

Ruth Levine: Thanks very much Lawrence and thanks to all of you for coming over on a hot day. I don't have a PowerPoint presentation, so this is the place to direct your attention. Some projects start with a brilliant vision and some projects start with some kind of deep insight, but the project that we'll talk to you about today actually started with frustration. It's frustration derived from seeing that we're actually not doing all we could to achieve the core goals of the development enterprise. But as you'll see, what started with frustration, I think led to something that I hope will give us some significant hope about doing better in the future. The frustration I'm referring to is one that I have personally felt over the years working in the development business, being asked to design, implement and advise developing country governments on complex development programs in the health and education sectors, with little more to go on than some theory, my own on-the-job experience and observations, some good advice from colleagues and the occasional case report in the literature. Millions of dollars and many lives have been at stake. How big should the secondary school scholarship program be in Tanzania? Should we introduce user fees in the Bolivian Health System? Though economists think that

that's going to increase the efficiency of health service delivery, but the public health specialists are worried about whether it will undermine use of basic health services. With the new, more expensive model of delivering maternal and child healthcare in Argentina, really better than the old cheaper one. In 1990, when I first started working in development, I was unable to find any systematic body of knowledge from the decades of experience implementing similar programs. The situation actually isn't much better today. $30 billion of development assistance are spent annually on social programs and many times that amount by developing country governments, on well-intentioned development programs that are based on very weak evidence about what works. So that was my frustration. But then there's the frustration we've heard time and again from high level policy makers and their staff who, with the stroke of a pen, can determine whether or not development assistance rises or stagnates. So these people, and I know that there are some of you here today, scour document from bilateral and multilateral development agencies, searching for clues about whether past appropriations actually made a difference in people's lives in developing countries. So instead of that evidence, here's what they found. They find statements like, "This program was highly successful in completing the activities anticipated in the program design; however, due to limitations in the data, lack of baseline survey, we don't know what the impact was on the target population." Well these are not findings that inspire the taxpayers of wealthy countries to open their wallets or to keep them open. And most important, there's the disappointment of policy makers who are genuinely trying to squeeze progress toward health and education goals out of very limited resources. So this is an excerpt from a letter I got, actually last week, from Timothy Thahane, who's the Minister of Finance in Lesotho, and he heard about this work and wrote, "There's increasing frustration among us in developing countries about the prescriptive fads regarding what works and what improves the quality of life for all people sustainably. Based on limited samples, incomplete and poorly formulated, analyzed or articulated empirical studies, fundamental and far-reaching economic and social and environmental policy recommendations are made to African countries and massive resources spent to support the implementation of poverty reduction at high consultant costs." He goes on to say something needs to be done about that. So in other words, while those who are footing the bill for development assistance start to ask hard questions, those on the other side of the table are asking even harder ones, and being unable to answer them after so many years and so many dollars, it's really a shame on us.

So about two years ago, deeply worried about this, Nancy Birdsall, President of the Center for Global Development, Bill Savedoff, a colleague who works at Social Insight and I decided to get a group of smart people who know a lot about development, together to think about what could be done. This was the Evaluation Gap Working Group, which was supported by grants to CGD from the Bill and Melinda Gates Foundation and the William and Flora Hewlett Foundation. So by looking at what different institutions are doing by analyzing the nature of bureaucracies in both developing countries and in rich countries, and by arguing a lot with one another, we came to a set of conclusions about the reasons for the shortage of knowledge about program effectiveness, and we developed recommendations about what might be done. Then we asked a lot of other smart people, and I really do mean a lot of other people, what they thought about our ideas. So we did this through a web-based survey and through consultations in Washington, California, Cape Town, Delhi and Mexico and London. We also got a constant stream of emails from angry people. Some were annoyed with us, but many were annoyed about this kind of perpetual blind spot in the development business. And all of that input we got helped us to refine

our arguments and our recommendations and to finalize the report that you have today. So in its simplest terms, the problem that we set about to understand is that those in the development business and decision makers in developing countries who care about improving health, education and other social outcomes, don't have the information they require to really make good decisions about how to spend money. Now, of course, we're not completely in the dark. But unfortunately, most of what we know is about the nature of social problems and the populations affected. We actually know a lot about how bad things are. We also know, particularly in the health sector, the effectiveness of particular interventions in highly controlled situations. So we know that immunization works, we know that oral rehydration therapy works. So that's good, but it's not enough. We have a reasonably good sense, or we think we do, of how improvement in particular types of social indicators would contribute to a broad and sustained economic development. There's been elaborate analytic work to show the impact on economic growth of improvements in health outcomes and education outcomes, female empowerment, access to credit and so forth. So that's also good, even if most of this work is motivated by an impulse to kind of make a point. But also not enough. The crucial piece we are missing is knowledge about how effective particular types of social programs are in changing social outcomes. We don't know, for the most part, what happens to school enrolment and drop out rates when we change school inputs or management. Is it textbooks, blackboards, smaller class sizes, more training for teachers? What can make a difference? We don't know the impact of introducing social insurance on health service utilization or health itself, even though such an insurance is usually justified in part by claiming it produces health benefits. So these aren't abstract academic questions. Answering them is really the only way to make smart choices about spending precious public and private resources. So we don't know about these things. How can we advise developing country governments how to spend money, their own, ours? How can we ask the U.S. and other rich country taxpayers to support a scale of development assistance? And certainly more important, how do we answer that Minister of Finance from Lesotho and the people he represents?

It's tempting to say we don't know what it takes to mount effective social programs because it's actually unknowable. We heard this many times during this work. Human behavior is so difficult to observe. Each context is so different. Making the right comparisons is so tough. Maybe it's unethical. It's just not doable. So we should stick to measuring things we're used to measuring – inputs, activities, outputs, dollars spent. Forget about impact. Well those arguments would hold a lot more water if we didn't have a growing set of examples from around the world in which program impact was measured and measuring it made a difference. In the U.S., for example, the evaluations of job training and income support programs have shown that large scale, complex demonstration projects can be credibly, rigorously evaluated in ways that have highly policy relevant results. In Mexico, a first rate evaluation of the impact of a conditional cash transfer program both confirmed that the program met most of its core goals and provided valuable information to improve the program. And we have examples on a smaller scale from Kenya, India, Indonesia. In short, it's not just doable, but it's done. The problem is that it hasn't been done in a systematic way at an adequate level. The few times when good evidence of program impact has been generated, have been the result typically of particular researchers' interests and opportunities. The kinds of studies from which we get information about program effectiveness, impact evaluations, estimate net impact by comparing the changes that occurred in key outcomes among program participants with the changes that might have

occurred among similar populations not served by the program. They address what many in this room call the counter factual. Conceptually, the most straightforward way to make the comparison is to randomly assign people to participate in the program, some people to participate in the program and others not to. It's actually feasible much more frequently than you might imagine and has been shown to be feasible when you're scaling up a program. When it isn't, there are alternatives – good design, careful design coupled with good statistical methods. The important thing to remember about impact evaluations is that the design and data collection for the evaluation has to be initiated at the time programs are started. You can't start the evaluation design after the implementation has been done. It's too late. It has to be part and parcel of the design. The methodologies to do these sorts of evaluations and the demand for the knowledge they produce have been around for a long time. In fact, a couple weekends ago, I had the pleasure of reading a book by Alice Rivlin from 1971 that laid out the entire argument. We've also heard for many years calls for more rigorous and independent evaluation of development programs, also not new. But outside of a few agencies, relatively little progress against this has been made. So the problem then starts to look not like a lack of the evaluation technology, but rather like a problem with institutional incentives. There are three basic incentive problems. First, a portion of the knowledge that's generated through impact evaluation is a public good. So the people who benefit from the knowledge include, but go far beyond, those who are directly involved in the implementation and funding of a program. So for example, when the government of Bangladesh rigorously evaluates a girls' scholarship program, that knowledge, yes, benefits Bangladesh, but policy makers in India, Pakistan, even Senegal can use that knowledge as a point of reference. These broad benefits are amplified greatly when the same types of programs are evaluated in multiple contexts, so then you really start to create a body of evidence. But the cost benefit calculation for any particular institution doesn't include those broad benefits and so, for an individual institution, it doesn't look worthwhile to do impact evaluation. The second incentive problem is that the rewards for institutions and importantly, for the professionals within them, come from doing, not from building evidence, not from learning. So those who work in ministries of education, ministries of social development, USAID, the World Bank, any number of institutions, are motivated to serve people by getting programs up and running. In fact, the measures of this sort of activity, the projects launched, the money spent have, for a long while, been just about the only things that we looked at as the performance of the institutions. So in this sort of environment, it's extremely difficult to protect the funding for good evaluation or to have the patience to wait for the evaluation design to be built into the program itself. There's an urgency to get started, forget about the baseline study, use the evaluation money for serving a few more people. Time and time again, the resources for evaluation are cannibalized for program implementation. And third, I think we have to be candid and acknowledge that there are disincentives to finding out the truth. If program managers or leaders of development institutions or ministers of social development think that their future funding is dependent on achieving a high and consistent level of success, then they're not going to be tempted to take the risk of doing an independent evaluation, which might show that some of the programs aren't working as well as they claimed. The temptation instead is to focus on producing and disseminating success stories. The concern over the consequences of unfavorable results, I think, is very deeply woven into the fabric of most bureaucracies. It's really a rare institution that's comfortable with independent, clear eyed, evaluation of where its investments have paid off and where they haven't that will share information in a transparent manner and then will make adjustments accordingly. This is particularly the case when peer institutions are

behaving similarly or worse. There's really no benefit to having a reputation as the institution that's best at learning from its mistakes, when one's competitors apparently don't make any.

So we have three big challenges. We've got to find out how to fund public goods. We've got to safeguard the funding for evaluation and we've got to reward honesty and learning. But I will have failed if I leave you today thinking about this as kind of a technical, bureaucratic problem with a technical, bureaucratic solution. You know, it really goes much, much deeper than that. I think the core of the problem is about our collective ethical position. If we fail to learn as we go and we adhere to a kind of arrogance of believing that we already know far more than we do. As Will Rogers put it, it isn't what we don't know that gives us trouble, it's what we know that ain't so. When the quality and quantity of people's live hang in the balance, and that really is the case for the kinds of programs that we're talking about, our willful refusal to learn, our refusal to learn how to make those programs work better, really becomes a serious moral problem, and it's time to solve it. Perhaps the most gratifying and inspiring part of this project has been hearing from many people who sometimes at some professional risk, have offered support for the ideas that we've been trying to talk about in this work, and are deeply committed to solving these problems against, I would say, sometimes quite steep odds. Now I know you're actually interested in the group's recommendations, so I'm very pleased to pass the baton to my friend and colleague, Bill Savedoff, who will share those.

William Savedoff: Thank you Ruth. I do have a few slides here, so I'll take attention from me and you can put it up there. I wanted to just emphasize that what Ruth has presented, this trajectory that we've made from frustration and deliberation and interviews and meetings, which are all documented at the end of the report, we have an appendix there with over 100 names of people we've talked to, and I want to really thank the people who volunteered their time to work in the Evaluation Gap Working Group, who spent a lot of time in email discussions, but also came to Washington several times to help us battle through all these ideas and try to clarify them and come up with something that we think will really tackle the problem. Because again, the idea here wasn't simply to say there isn't enough being done. The idea was to say, why do we consistently come back to the same question, why isn't more done?

So if I can have the next slide. I wanted to start by reiterating some of the points Ruth made about what is a good impact evaluation, and we did see a number of cases and look at why they were successful, and these are the main characteristics that we saw. The first thing is that these evaluations tended to start in the design phase of the program itself, largely because ex ante, there was some effort made to think through what would be an appropriate way to get valid inferences? How would we define the counter factual? What's an appropriate control group? Secondly, these impact evaluations, unlike the sort of the model of evaluators as coming in objectively from the clouds and sort of looking down and seeing what's going on, actually the most successful impact evaluations we found involved the policy makers and the managers from the start. And part of this is because those were the evaluations that actually looked at questions that were relevant to the policy makers and managers, and partly because they had a buy in and so the data got collected properly and there was follow through. And here I want to mention that, although this initiative started really within CGD's theme of aid effectiveness or looking at donor monies, it became very quickly apparent that the donor money really isn't the key issue here, that impact evaluations, if they have a value, it's to changing policy in developing countries

themselves. It would be great if the donor money were also spent more effectively, but really, the key stakeholders here are the policy makers like this finance minister in Lesotho, who are thinking about their own country's very scarce taxpayer money. The third thing is that they include external actors and external actors help with expertise with quality and with integrity. They're the check that allows the study to move forward without watering down its conclusions or subtly massaging the results. It was very important, for example, in Progresso, that the academics were involved, an external agent. It gave, the people in Mexico said it gave the domestic audience a sense that this was a credible external study or externally validated study, that it wouldn't have had if it had just come out of the government.

The fourth thing, and this may sound kind of silly, but a good impact evaluation has to focus on impact. There are a lot of studies that are called impact evaluations that contain a lot of very valuable information about processes, institutions, operations, inputs, outputs, efficiency of the program itself, implementation, but they don't measure impact. They don't attribute the outcomes that you observe in the population to the specific program.

The fifth characteristic is they have to document the context and the process in the operations. You can't forget that information. That's important to improve the validity of a study to test whether the underlying model actually makes sense and it also helps later on when you're trying to generalize the results of the study and say, "Would this be relevant in another country?"

Sixth, they have to address enduring questions. Impact evaluations do take time. You can't start one and expect results two months later, except in very rare circumstances I imagine. So we want to be addressing questions that are really of longstanding importance, and I think it's really dramatic that you still find that the Rand Health Insurance studies of the 1970s continue to be referred to in literature because they were very substantial, solid results. They told us some very fundamental things and gave us a point of reference, and are still talked about today, because they were looking at a question that's of enduring importance.

And finally, good impact evaluations have to be selected and strategic. These are not studies that we want to do on every developing country program or project or donor program. The idea, the value of these studies is the knowledge that comes out of them and we want to do these studies on the programs that will tell us the most, meaning they're addressing big issues, that they are programs that can be evaluated, that can give us valid information and preferably, programs that have some generalizability, that it's a context that isn't so unique that it doesn't tell us anything that's relevant to other places.

Now that's what the good cases were. But the frustrations that Ruth was talking about come, I think, the way I've put it here, is separation of three different people, three different actors in this process – the researchers, the managers and the policy makers. In my work at the Inter-American Development Bank, where I first entered this world, I was most familiar with what was happening with the researchers. Looking at the studies that were supposed to guide me in helping design a new program, so often I came across a paragraph that said, "We want to measure impact but oops, there was no baseline data." Or "Oops, there was no unbiased control group." Or "Oops, we didn't collect the outcome data." It's embarrassingly common. The other side is for the managers, the project manager. When there is an evaluation designed into a

program and there's money there, the managers were focused on implementation, and that's what they were going to be held accountable for. And so, whether it was lack of money or lack of time or attention, it was too common to find three years into the program, that somebody said, "Oops, we were supposed to do a baseline survey." And finally, the policy makers who, all too often, come to the research department or the evaluation department and say, "Well, we'll give you two months if you can just pull together everything that we know about preventing HIV, teacher training and so forth." And then somebody has to do their best to find what's there, to say, "There's only two studies that were really done with valid results, but we think we can extract the following lessons from the other 50 studies." If you want to see some examples of those, the appendix, again, we put a somewhat amusing collection of them into the appendix of the report.

Now what the Evaluation Gap Working Group tried to do was look at the incentive issues, and this is the way we think it could be. This is the crux of what needs to happen if we're going to make a really serious change and not be returning to business as usual. And it really requires countries, again I want to emphasize, we're talking about developing country governments themselves, as well as the agencies, thinking in these terms. First, splitting the impact evaluation process from program approval and implementation so that it can be selective, as we were mentioning earlier, so you can concentrate financial and technical resources. It isn't as useful to do 1 percent of each project dedicated to impact evaluation, when it might be better to do 200 percent of three projects on impact evaluation. Three good teacher training impact evaluations would have leverage over what's done and designed into hundreds of other programs around the world. And finally, by splitting it off, making it more independent and credible. But paradoxically, we also have to link the impact evaluations closer to the programs in the design phase as I mentioned earlier, so that the questions will be relevant, so that data collection will be appropriate, and so that the conclusions can be rigorous.

So the Evaluation Gap Working Group came to two large recommendations. The first one is that ministries, agencies and NGOs should do more of what they do, what they can do in terms of impact evaluation, and more of what they are doing. There are a lot of initiatives out there, the development assistance committee, evaluation network, is doing a lot of work trying to improve evaluation. The ECG with the development banks, numerous bilateral agencies and multilateral agencies are trying to improve training in capacity to do evaluation work. Countries like Mexico actually passed legislation requiring impact evaluations. There are NGOs and research centers that are pushing this agenda forwards. Some of these things are complementary to impact evaluation, they're not actually impact evaluation itself. And most of the evaluation departments in donor agencies and development banks are working very hard to provide very useful information about processes, operations, strategies, country portfolio performances and monitoring what is being done. There's a lot of data collection going on. There's capacity building. All of these are complementary and necessary and need to be strengthened to continue if impact evaluations, the information from impact evaluations is going to be useful in context. There are also a variety of initiatives specifically aimed at improving impact evaluation. The World Bank has the dime initiative, the French Development Agency just recently contracted a very interesting impact evaluation on micro finance. I can't list them all right now, but there are individual efforts that are going on and they need to continue, they need to be strengthened. But we really, as a group, don't think that we're going to solve the fundamental problem, make a

leap, know that nine years from now when we look back at how we've done with money spent for the MDGs, feel like okay, we may not have reached the goals or all the goals, but what did we learn in the process, unless there's really a collective effort to reverse or to address the incentives that impede good impact evaluation. And these are things that really can only be done collectively. Collectively, developing country governments, agencies can identify those enduring questions by discussing with each other and coming to a set of priorities – where's the money going? What are the big problems? What are we going to want to know five years from now? The second thing is promoting strong standards. It's much easier to hold each other accountable, to raising the standards, than for an individual organization to maintain the process of raising standards within itself. The third thing that a collective initiative can do is give some independence to the review process. And that also strengthens the backbone of the process. Several, many people through this process, mention that there are often very small windows of opportunity when an impact evaluation can be designed. When the right constellation of factors, an interested project manager, an interesting question, a country that's eager to do this, an outside agency that might put in money for an impact evaluation, and what's missing is a little bit of money to bring in a technical expert to find out, is this valuable? Can we evaluate it? Would it be worthwhile? How do we design it into it and make it happen at that moment? So some targeted assistance, very small money, can be very catalytic. And finally, there have to be funds, substantial funds dedicated to good impact evaluation in some form or other.

So, rather than simply laying out, you know, this is what has to happen, we tried to move in the direction of some fairly specific ideas and the Evaluation Gap Working Group, the report you'll see, lists a series of functions that we feel need to be done collectively. We prioritize the ones that we really feel are core functions that are both agreeable and decidedly collective in nature, and a series of other functions that an initiative, a collective initiative, could undertake. We looked at institutional options and their pros and cons, how they would work, and some funding options, and we're placing those out before a group of people in Bellagio, Italy next week and before other agencies and groups to see whether there's some buy in for one or more of these ideas. What I'll put up here now is the broad pattern of what we're talking about, because it should give an idea of how this would address the incentive problems. The first thing I want to emphasize is on the left hand side. The developing country governments, development agencies and implementing NGOs are the key. The rest of this is serving them and currently, the demand for better evidence and the supply of better evidence is just stuck within that circle. So the idea of this initiative is to break it out so that the demand can be articulated in terms of the enduring questions that are being asked and the kinds of programs we want to see evaluated. Members of this initiative, which will be voluntary groups, organizations, governments that want to be pioneers in pushing this forward, who would provide funding or dedicate their own funding for impact evaluations internally and hold this initiative accountable. We're calling it an International Council on Impact Evaluation – that's the term we're using – that would manage this process or would serve the members in implementing the kinds of functions that are agreed upon. The council would be linked with experts who would provide the panel reviews and the expertise for designing impact evaluations and keep the standards and rigor up to snuff. Very key in the low right hand corner, the three categories of people who have to talk together at the beginning when an impact evaluation is designed and carry it through, and the council could provide them with networking, technical support, with quality promotion and channeling funding

from members if that's decided upon. And ultimately, if this works, there would be a more systematic and steady supply of good evidence going back to the main stakeholders.

So to conclude, I'll just say the imperative is to do something. That's what frustration and urgency sort of says. Everyone in this process has been fascinating. Everyone recognizes it's a problem, but there's substantial inertia. So the number of conversations we've had with, "I'm so glad you're working on this, but gee, that doesn't sound like we could do that." Or, "Yes, we need to dedicate funds, but I'm sorry we don't have any." Or, "That's a great idea but we don't want to hold ourselves accountable to somebody else's standards." It comes down to that question, "Who will bell the cat?" And that's again why we really feel it's a collective initiative. We have to get pioneers who are willing to take the risk and say, "Look guys, we're going to go out on a limb, but we'll do it together." And what's exciting at this moment is there really is a confluence of interest, urgency and support from major foundations, from several developing country governments and several agencies that might actually make this a reality. So I'll stop there. Thank you very much.

Next Speaker: Bill and Ruth, thank you so much. Right now, I'd like to invite our panel to join us up here. Do be careful as you climb up on the stage, and I don't know if the – I guess we may have to put up with the screen. We normally have this very cool screen that comes down from the ceiling, but it's broken today, so – the screen's going away?

Next Speaker: Yeah.

Next Speaker: Okay. So just hold off for a minute and the screen is going to be removed. And we'll get that light out of your eyes. We have an extremely eminent panel today. Whenever I moderate one of these panels for CGD and I read people's bios, I always think, "Well gee, what have I been doing with my life?" And if I had to write down my list of achievements it would look pretty pitiful compared to these. I'll start introducing our panelists as they're taking their seats. To my immediate left will be Jon Baron. He's the Executive Director of the Coalition for Evidence-Based Policy. John founded the Coalition and serves as its Executive Director. Based on the work that he did there, he was nominated by the President and confirmed by the Senate in 2004, to serve on the National Board for Educational Sciences, which helps set the research priorities and agenda for the U.S. Education Department. Next to Jon will be David Gootnick, the Director for International Affairs and Trade at the Government Accountability Office, the GAO. David is also a physician and, among other things, was a volunteer in Malawi, working against polio, early in his career. In the center will be Kenneth Peel, the Deputy Assistant Secretary for Development, Finance and Debt at the U.S. Treasury. Kenneth, or is it Ken?

Next Speaker: Ken.

Next Speaker: Oversees U.S. participation in multilateral development institutions including the World Bank, the IFC and the regional development banks, the Global Environment Fund and the International Fund for Agriculture Development. To Ken's left of Nilmini Rubin. She's a Professional Staff Member for International Economics in the U.S. Senate Foreign Relations Committee. Among other things, she drafted Senator Richard Lugar's Multilateral Development Bank Reform Legislation that became law in November 2005. Nilmini, we're delighted to have

you with us today. And finally, Franck Wiebe, Managing Director for Economic Analysis at the Millennium Challenge Corp. Franck is very bold. He's been in this job for is it – a few days?

Next Speaker: About ten.

Next Speaker: About ten days. He's agreed to come, nonetheless. Before he joined the MCC, he was the Chief Economist at the Asia Foundation where, among other things, he was responsible for economic reform and development programs of the foundation throughout Asia. I'd like to start out with you if I may, Ken.

Next Speaker: Great.

Next Speaker: I previously worked at the World Bank and was occasionally was on the receiving end of missives from Treasury – I should say my bosses were – saying, "Can you get us some evidence that what the bank is doing works?" So I imagine that you or your colleagues or your predecessors might have been the people who originated those missives, and I wondered if you could say something about your experience in looking for evidence of the impact of the multilateral institutions and how that might connect with this initiative that we're launching today.

Next Speaker: Okay, well thanks. First of all, I'm going to take a little detour. I'm disappointed to find that I'm not the new kid on the block. I've been on this job, I guess about 3 ½, 4 months now. But, you know, I have been around Washington 20 some years. I've been a federal government employee for 22 years and I haven't spent a single day in any civil service, any foreign service, any military service or any accepted service, so if I figure out how to do that, I'll let you know later. I spent the last 3 ½ years working out of the White House, where I worked with Nomini for a while on the NSC doing international, environment and energy work, so I'm going to duck your question to a certain extent, but then come at it from a little bit of a different perspective because you know, you do get right at the heart of the matter because that is – anyone who's been in the Executive Branch, at some point and often frequently, gets those exact calls from ONB, you know, when they're evaluating your programs and your policies and they want to know, you know, as they're looking at the next budget, well you're asking for the same amount or maybe you're asking for more for this program – do you have any evidence that it's actually doing anything? Is it actually accomplishing anything? And it's, you know, it's sad but it's humorous in its own way. I mean I think we've all seen evaluation write ups of an agency and say, you know, yes, we achieved all of our goals. We spent all of our money. Well no, spending your money is not your goal – actually achieving something is your goal, and the amount of evidence that you often have for that is, you know, sadly very thin. And you know, this is a phenomenon, not just in the, you know, 150 function for those, you know, who spend way too much time in Washington, in the foreign affairs function or in the development function. It really, it is, this is a problem that is not, again, is not just are your development monies being spent correctly, but is your government money being spent correctly? Or is your foundation money being spent effectively? And it impacts, it has an impact on all sorts of government programs, whether domestic programs in the developed countries or, as you point out, your own domestic funding in developing countries. You know, [unintelligible] afford it the least. So I don't think I've answered your question, but you know –

Next Speaker: I cut you off because I want to set a good example for other panelists, that when I said brief answers, I meant it. So I'm going to come back to you and give you another opportunity, but now I'm going to Nilmini and, Nilmini, among the things that you have been, I think, very active in pursuing in your role in the Senate Foreign Relations Committee is concerned about corruption in international development issues. And I wonder if you have come across or noticed a dearth of sound evidence about what works in the fight against corruption? I'm a little concerned that we're charging off on a crusade that, once again, we don't really know what works. Is that, do you feel that that's true in corruption, or is corruption an exception to this rule?

Next Speaker: I think corruption's consistent with the other sectors, health, education. There is a dearth of knowledge on what has worked in corruption and what hasn't. I think Senator Lugar has focused on what the basics of the anti-corruption effort is, which is making sure we know where our money is going, you know, just kind of the first level questions. But, kind of going past it and thinking about what methodologies in projects and what pieces that are built into the projects actually work, I don't think there's any better study of that than of any other issue.

Next Speaker: Thanks very much. David, I was struck when I was looking at the bios to see that you are trained as a physician, and among other things, served as the Director of the Office of Medical Services at the Peace Corps. And one of the things, when I've been discussing this with Ruth, that always strikes me, is the insistence and assumption within the medical community that you would have controlled trials. You wouldn't approve anything if it hadn't been subject to a controlled trial, and I wondered if you're straddling both the policy and development realm and medical science, and can share with us, you know, why this has happened in medicine and what the potentials are for applying some of those lessons in social policy.

Next Speaker: Right. Well thanks for the question and, yes, I am a physician by training although, having spent the last ten or more years in government, I often refer to myself as a recovering physician, but it kind of begs the question as to, "Recovering from what?" So I won't go too far into that. But the bio medical model tends to use randomized, controlled, double-blinded, placebo-controlled studies as kind of a gold standard, and uses the notion of a discreet intervention in a specific patient, looking at results in the randomized double-blind, placebo-controlled prospective manner, is the gold standard for documenting results. I think there's some similarities and some distinctions. It's a great notion to bring that level of rigor and that kind of standard to bear on social sciences research and certainly the challenge of outcomes, research in the 150 account or in the developing world setting. And the remarks from the document really do speak to both the notion of prospective certainly. Randomized, yes, although there are some caveats. When you look at the Mexico study, how was that randomized and, to the extent that randomization really does achieve controls, is a greater challenge here. The notion of control that both tests intervention versus placebo and controls for everything else, either through the collection of data or through the analysis of data, typically using regression to determine that really other things are held constant. All three of those, to a greater or lesser extent, can be handled. I think the notion of the double blindedness is really where it goes off the cliff a little bit with respect to the biomedical model, so that, does the observer and does the observed have true independence there? Is there a Hawthorne effect? Is there a placebo effect? There's an

interesting phenomenon in the medical literature, the nocebo effect, which is the inverse to the placebo effect. If you give patients a pill and say, "This is really useful for your pneumonia but it may make you itch", more of them will itch than if you just say, "This is really good for your pneumonia" and go on your way. So the observer bias is important. I think the Mexico study is an interesting case about whether or not the observer and the observed bias are removed. And then finally, perhaps the more realistic analogy is the notion of operations evaluation research in the medical arena, where the Rand study was mentioned. That's really health services research, but kind of cues us that, both with respect to the time that's required and the outcomes that you can achieve, you really better be patient, that the notion of a global public good here is, I think, really key because these things take an enormous amount of time. What you're going to be able to get out of them is incremental and not transformational. Rand was really interesting. It started in the 70s when Nixon had a universal healthcare proposal on the table, and ended in the 80s when Regan was implementing DRG. So the fact that it showed financial barriers to care caused less consumption of care, was used very different than it was initially designed, than the purpose for which it was initially designed, and Rand too, which ultimately looked at health outcomes, had a very limited set of observations for the poor who were hypertensive or for the poor who needed vision correction, lack of health insurance made a difference in health outcome, but you really had to tone down your expectations as to what you'd get.

Next Speaker: David, thank you. Jon, you worked on evidence-based policy, I gather, primarily in a domestic context, is that right?

Next Speaker: That's right.

Next Speaker: Could you tell us about some of the experiences there and, I mean, obviously, evidence-based policy is in great need everywhere, but I gather that one of the reasons this initiative, our initiative if focused on development is, aside from the fact that that's our main area of concern, is that the dearth there is greater and therefore, I am presuming that in domestic policy, it's somewhat ahead. Can you share with us some of your experiences there, your perceptions of where it's working and maybe where it isn't.

Next Speaker: Yes, well, I think that the problems in domestic social policy may not be quite as severe. The dearth may not be quite as large, but they're still very large. I think this report lays its finger on a critical problem in a number of different fields. Both development assistance, as well as domestic policy, which is that most of the programs, we work primarily in social programs in the United States, most programs – domestic policy, foreign aid and so on – are operating in a vacuum of scientifically valid knowledge about which things they could fund, do fund and so on, work and don't work. And I think the report lays that extremely well and we have also presented similar evidence and are seeking to address similar problems in domestic policy. Let me just go one step further, which is to say that if development policy is similar to other fields, there is very good reason to believe that most of the programs and interventions that are ongoing, or at least many of them, if they were evaluated in a truly rigorous study, a well designed, randomized, controlled trial, would probably be found not to work or to be marginally effective. And I say that, in domestic U.S. education policy, when good randomized controlled trials have been done, they typically find that a whole lot of things that everybody thought were going to work, that were good ideas that were backed by less rigorous studies, that many of those

don't work. A few work, but many of them don't. Just a couple examples, a randomized trial that was done of the federal government's dropout prevention programs, early literacy programs, the big after school programs, HHS's comprehensive child development program. The same is true in medicine, where there are a lot of interventions that do work, but that's because many, many, many more were studied a whole lot of those were found not to work. Hormone replacement therapy for post-menopausal women, a lot of the dietary interventions. You know, we've all been hearing about, we've all been told you should reduce fat in your diet and increase fiber and vegetables as a way of prevent colo-rectal cancer. And that's what all of the correlational studies and the cross country studies and the migration studies have shown. But one of the clear findings from the large women's health initiative, the randomized controlled trial that the NIH sponsored, there were some ambiguous findings from that. One of the clear findings was that that kind of dietary intervention has no effect, no effect on colo-rectal cancer. So I would just sort of leave it at that. Many things don't work and I'm going to stop there, but I'm going to say the trick, I think, is to do what has been advocated here – to get a lot of these studies underway of a lot of different creative approaches, so you can identify a few that do work. In domestic policy, there are few examples, maybe ten examples, of interventions that have been shown highly effective in well designed randomized controlled trials. The trick is finding that goal amidst everything else that's going on.

Next Speaker: I guess if most of what we're doing doesn't work, once we learn that, there'll be lots of resources available to do the things that might work. Frank, you've been patient. I'm tempted to ask you about your view from the Millennium Challenge Corp, even though you've been there only recently, because the Millennium Challenge Corp is, I think, unique among U.S. foreign assistance programs, set out to really be run along rigorous business-like lines, and a portion of that is doing things that work, and I wondered if, since you'd been there and talking with your colleagues, you have a sense that they have the evidence they need or not. And I'm going to give you a dodge as well, which is, feel free to draw upon your experience with the Asia Foundation in providing your answer.

Next Speaker: Right. Speaking from my wealth of experience at the MCC, one thing that struck me in reading the report prior to this seminar today, was how similar the recommendation in the report sound to me like the work plan that is being implemented at the MCC. The resources being dedicated to monitoring evaluation go along with each activity the MCC funds, so every project does baseline research and monitors implementation. Now clearly every activity can't be evaluated, just like the report advocates, only selectively will activities really be probing challenging questions. And the MCC has set aside resources to do that as well and so, as we roll out, I think that we'll be seeing MCC implementation that very closely follows many of the recommendations that are laid out here in the seminar, in the report. So, you know, taking the risk of sounding like an organizational [unintelligible] on my first public appearance, I think that that should be noted, that the institution has learned from some of the mistakes that other institutions have made and are trying to build, the institution is trying to build on that. I think it's important to recognize that the MCC is also in a unique position as an institution. We're making grants. The partners that we're working with want to get those resources and are willing to agree to allow these activities to be studied in a rigorous and independent fashion. This is being planned up front and so is part of the agreement, again, as the report recommends. Most institutions aren't in that position. There are negotiations that take place. The political process

through which most institutions work is a lot messier, a lot more complicated and so, while on the one hand, I look at the report and I say the recommendations make a lot of sense. I'm pleased to be working in an institution that can work in this fashion. I think we shouldn't be, we shouldn't be naïve about our prospects for doing more of this under other circumstances. I think the disincentives, as the report identifies, the disincentives throughout institutions to allowing independent observation and analysis, are very, very strong. Many activities go forward despite best thinking on theory, despite what people know hasn't worked in the past. In those kinds of contexts, why do we think that more evidence is going to change the practices? That doesn't mean it's wrong to do that. It means we need to go in with our eyes open about what we invest in that and recognize that there's a lot of better policy, a lot of better project design that could be done if it was based on best knowledge as it stands now.

Next Speaker: Franck, you've hit on something that I think is really at the core of this report's two recommendations. The first recommendation is that those who are currently involved in evaluation, those agencies and individuals and researchers that are doing it, should keep on and do better and do more. But that's not really a recommendation for change so much as an exhortation. The core thing that's new to my mind is the creation of an independent entity to address the public goods nature of the problem, to express a collective desire to address this solution. And so I'd like to ask the panel, each of you, your views on whether or not there is in fact such a need or whether such an independent entity could help to address the problem and Ken, since I rather rudely cut you off –

Next Speaker: No, that's fine.

Next Speaker: I'm going to ask you to go first. But then I'm going to sort of let people volunteer in order as you would like to address this. The core question of whether such an independent entity makes sense and can help solve the problem.

Next Speaker: Let me first say that I think the report is, we agree with – let me just talk about myself – I agree with essentially almost everything in here, with the one exception of this idea. And the reason for this is that I think there's actually more consensus on the importance of evaluation work than it sets up to be. Much of this report is making the case for why this work is so important and I think that there is a lot that's beginning to get underway and you touch on that in the paper, although you don't go into that as much as I certainly was hoping to see. But I think the danger of trying to set up an independent separate sort of arm's length new international institution of this sort is, I think is the danger of ghettoizing the effort. It's almost, it reads as too simple a solution to a real serious and, I think, evident problem and I think it seem sot be too simple a solution because I fear that it is too simple a solution. I would, you know, I would recommend, and we would recommend, more of, like for instance on page 30, you summarize much of what's already going underway and holding those kinds of efforts together and making them more systematic, I think, is a good idea. And there's much going on at MDBs, the MDBs in this are much more active than in the past. They need to be a lot more active than they are right now. But I'll just leave it at that. There's much more to say.

Next Speaker: Thank you very much. I'm happy to start choosing people, but if somebody - no volunteers? David?

Next Speaker: Speaking in my personal capacity, you can scratch GAO off my nametag here – I thought it was a refreshing idea. I thought it was interesting. I thought it was a novel and, feasibility aside, I thought there was a lot of good rationale for it. In particular, the issue of the notion of independence and the notion, as Ruth pointed out, that in the face of statements that this stuff is not, is just not doable, the point was made that it is doable, but it's really hard. It's hard, it's time consuming, it's expensive and you need the luxury of time and the luxury of somewhat diminished expectations about what this is going to show. It's not the keys to the kingdom, and the observations that will be made five years down the road will say that in this time, in this place, under these circumstances, these economic conditions, etc., that this intervention worked or didn't work. You can bet MCC is not working towards fulfilling the no hypothesis. They are very invested in showing that their programs work. And they're not going to be happy coming back to Nilmini five years from now saying, "Hey, we've just found that this road did not achieve the rate of return that we expected to do, and here's our budget request." It's just not, it's just not really a practical hypothesis. The other thing to say about MCC, with all due respect and not to sound like an auditor for the moment, but MCC should be commended for spending a lot of resources on outcomes type research but, in reality, what they are doing, their signature outcome metric is economic rate of return, and it was very nicely laid out in the report the different forms of evaluation that can be done, and that a cost-benefit analysis, they're doing a form of sort of a classical cost-benefit analysis. For every dollar they put in a project, they seek to get dollar 20 back out in economic benefits, and that will, in turn, that degree of economic growth will in turn be reflected in poverty reduction. It's a noble effort and it's a big challenge, but it isn't even at the level of the type of outcome prospective. An analogy would be MCC funds a road and they select a region where there is no road, that is truly the equivalent of a randomization or truly controlled experiment, and look at economic growth from one end to the other. They have, in fairness, built some IQCs, some contracts with US based firms to try to look at evaluation research. I'm not sure they've fully decided how they're going to use that money or that opportunity to contract with experts to do that and, even at that, I'm not convinced that those folks who are working for MCC are really testing the no hypothesis so much as they'd like to show, MCC, the folks who are paying the bills, that these programs work. So independence is critical.

Next Speaker: We'll give you a chance, Franck. You're wondering if we'll let you come back on that. Jon, you signaled me that you'd like to speak next.

Next Speaker: Yes. I think the report lays out a fairly compelling rationale for this sort of, this council, this central organization and I think a lot will depend, whether it achieves those goals or not, I think would depend a lot on the specifics of how it's carried out, which is sort of, you know, an obvious thing. But I think there is, and this is true in domestic policy evaluation as well, a compelling need for organizations to share knowledge about how to get rigorous impact evaluations underway and to use their results. There are a number of creative things that have been tried in domestic policy and also, other organizations like the Millennium Challenge Corporation, are now pioneering, to build rigorous impact evaluations including well designed randomized controlled trials into their grand programs or their assistance programs. And some of those things work and some of those things may not work, but there's definitely a need for an efficient means of sharing all the creative approaches to see which of those are effective or not.

Let me just give you one example from domestic policy, which I think may have examples for development policy, may be important. It's certainly important in a number of fields of American social policy. And that's this. This is the one area where I disagree with what David said. How's that for vague? Which is, I think that doing this may be easier, getting rigorous impact evaluations underway, may be easier in some cases than you might think. For example, the superintendent of schools in Seminole County, Florida decided he wanted to know which remedial programs for 9th and 10th graders were effective in improving reading achievement? So they randomly, the superintendent, the policy maker, was the leader behind this study and got a high quality evaluation team to work with him. They randomly assigned about 1500 struggling readers to one of three different interventions, reading different remedial programs, or to a control group. They provided the intervention, which they were going to do anyway, and they are testing outcomes at the end of the school year, that same school year, and at the end of the next school year, using Florida State Assessment of Reading Achievement. So the outcomes are being collected anyway by the state and for all the students, so there's no need for follow up. That whole study, start to finish, cost less than $50,000.00. It's a large, well designed, randomized, controlled trial. Large sample, 1500 students. It's not going to tell you everything, but it's going to provide a scientifically valid answer to whether these programs, any of these programs, is more effective than what the school was doing anyway. And it's an answer that will allow the superintendent and others to learn. There are probably a number of, you know, relatively low cost practical approaches to getting rigorous impact evaluations underway, and if you can share examples like that or things that the Millennium Challenge Corporation, or the U.S. Department of Education are doing to try to build, creative ways of building rigorous evaluations into their programs, I think some sort of organization to help share that could be extremely important.

Next Speaker: Nilmini, I'm going to call on you next because Franck's remarks about independence really led us into this round, and then I'll come back to you at the end Franck, for your views, not only about comments on the MCC, but more broadly about the role of an independent organization to address this problem. Nilmini?

Next Speaker: Well over the last 2 ½ years, Senator Lugar has had five hearings related to corruption of the development banks and, in addition, staff project visits and countless meetings. In the legislation that became law last November, it includes a little bit on evaluations, not as much as we'd initially hoped, but it does call for an independent evaluations office at each of the development banks, because we saw the need for more independence, not taking the further step that's called for in the report, but at least within the bank, something that's free from interference from management. And then at the most recent hearing where Ruth Levine testified, along with Adam Lerrick and Bill Easterly, on our second panel, the Senator asked us, the staff, to really think about what additional legislation would cause the actions we need, and one of the topics he wanted us to look at was the idea that Ruth brought up about the independent evaluation office, the arm's length entity. Because we do see many of the issues that Ken mentioned, you know, the idea of ghettoizing it or making it separate or somehow having it turn into something we never intended it to be. But we also share a concern that I think the authors had of the evaluations being corrupted by the organization, so that they're actually not being fair evaluations. And if you think about a bank or an aid agency, there's a person doing an evaluation and they say the power project was horrible, then they go back to the power office, I

can't imagine that will earn them a promotion. So we also need to be careful with how it's structured right now by keeping everything in house, and that, like the authors note, there could be real economies of scale by having all of us grouped, evaluations together, because baseline data is important and expensive. So there's definitely, I think, at least in our mind on the committee, some room to explore how to do it and we're currently struggling with how to be constructive in pushing it forward.

Next Speaker: Thanks very much. Franck?

Next Speaker: Well let me just start by saying that if all criticisms that I receive for my work at the MCC involved the praise of it being a noble endeavor, then I will feel happy about that. You know, I would just make two comments though in response to you, David. One is that not everything the MCC is doing is road construction. There is provision for funding both health and education programs and other social programs that are deemed to be expected to have returns that justify that investment. And so those projects will be relevant topics to be evaluated as they're implemented, and I think that the roads projects can be evaluated too. I think that, you know, I think your comment about not testing the no hypothesis is one that basically every institution that implements projects is vulnerable to. The fact that within the MCC, we have a separate division that is named Accountability, that we have within that a separate division that is monitoring evaluation, that is separate from the people implementing the projects, is at least an effort to try to address that, just in the way that what Nilmini was talking about in terms of having independent institutions within the multilateral development banks. So I think that it's fair to say that any institution that seeks to evaluate the work that it does is then vulnerable when it finds success to be, you know, to be criticized for it being, you know, somehow not independent enough. In this case, the MCC will also be doing contracts with independent institutions. Well then you say, well maybe they won't get another contract if they don't, you know, if they produce negative reports. I mean, again, we can get into a cascading process that probably even this international council, you know, at some level, might be vulnerable as well to, you know, if it receives funds from agencies that want that research done. So, you know, let's step back and say, you know, let's do the best job that we can and recognize that there are efforts underway that are seeking to do this and seeking to preserve objectivity and to learn from those lessons. I think that's what we're trying to do at the MCC. Let me just go then and turn to the report, and I guess my question on this is, it's just not clear to me how this international council, let me stick with the acronym that I know, the ICIE –

Next Speaker: Is that the Icky?

Next Speaker: My children would know it as an Icy and if they heard it described that way, I'd say nothing wrong about it. No, but how could it be different from what we have now? The donors, presumably the ones with money that are implementing projects, would be contributing to the ICIE, and then would be going to it for the technical expertise, the independence that it needs to do the evaluations. Well I mean, most of those donors have the ability to do that now, either in house or independently or through the implementing agencies. You know, it really begins to sound to me like the main difference is that through this council, there becomes a momentum from the outside, and external, moral sense of outrage for institutions that aren't doing that. And I'm just not sure I see that that's going to be the factor that tips it. In fact, you

know, in the report, we heard that many institutions are already doing more of this. The World Bank is doing this, USAID, MCC. I'm sure the Europeans are doing this as well. You know, there may be other ways to exert moral pressure on agencies to use their resources because again, these institutions have the resources that they could dedicate to this. They're smart enough. They could go to, they could go to the people who wrote this report, they could go to J. Powell, they could go the other places where the technical expertise currently resides and, through those contracts, those institutions would build capacity. And let's be very clear, that the technical capacity to do the kind of research that passes muster here is very thin worldwide. I mean very thin. It may be already that we're at the maximum that those people are doing. You know, that it would take a while for the interest we currently see, for the capacity to do really rigorous impact evaluation, to grow along with that market demand if you will. You know, so I guess what I'm saying is the motivations are clearly very, you know, are commendable. The notion that there's this, that there's this broad support to have more of this impact evaluation done, I think, is right. I think, at the same time, we could all exert moral influence on those institutions when you see a project being implemented that doesn't make sense from the beginning. Do we have to wait for an impact evaluation? I mean, again, many projects are being done over and over again, and we know they don't work. You know, especially on the economic development side. You know, I could, there are lots of thing that haven't worked, country after country, and within years, the donors are funding them again and again. You know, the agencies that know that have no incentive to turn around and tell them to stop because they're giving them money to do that. So I mean, I think that that notion of how we get kind of from a group like this, the moral imperative to the donor agencies to insist on evidence-based lending if you will, the policy making in countries is always going to be political. I think that that's the main thing, you know, the policy makers, cabinet officials, often know better than what they do because policies are being driven by local politics, just as they are in countries closer to home too. And so, you know, the only point that I'm making is I think that we're working in the right direction. I think that the idea of identifying who can do this work and encouraging people to use them and to do more of that work is right. I think the notion of building capacity is right and so, you know, let me end with all of those positive comments.

Next Speaker: I'm going to open this to questions and comments from the floor in a few minutes and ask Ruth and Bill to join the panel on the stage. Before I do that, I'm going to step out of my moderator role a little bit because I'm also a development professional on the communications side, and share with you my experience of watching independent evaluation being done in the bank. It's something that used to be called the Operations Evaluation Department, and it was deemed to not sound suitably independent enough and is now the Independent Evaluation Department, a sure sign if there was ever one, that there is a problem there, when you start changing the names of things. The number of times that I was involved, and I mostly worked in the Research Department, which was separate from OED, but nonetheless not responsible for the outcomes and the operations, but the number of times that I've been involved in communication planning meetings, about how to spin or bury the results of an OED report and hallway conversations about my God, it's coming up and it's not looking good, and then my friends who worked in OED.

First Speaker: But again, there wouldn't be an incentive to do that.

Moderator: I'm glad you mentioned that, because I find that it's the evaluators often, the in-house evaluators who are most nervous about this. But when I mention this to some other colleagues at the bank, again on the communications' side, one friend said, "Oh, that's great! We can outsource the evaluation." As far as she was concerned, it was you know one more thing we could get outside the institution. So, certainly going outside, if there were a credible place to go, I think would be a possibility.

Could you help me bring the chairs up? I'm pleased to say that we have near—we have a little more than a half an hour. We often find ourselves squeezing the few minutes for questions and comments from the audience. I'd like to give –I'm going to squeeze the panel a little bit. So you scrunch your chairs over and I'll keep trying from falling off the end here.

Next Speaker: I had just one brief comment. I found again much of this report really refreshing and quite even delightful in its own way. Talking about the problems about you know poor methodologies. On page 15, the one about self selection reminded me of a—of an article I read once about how red cars that—an analysis shows that red cars are inherently less safe than every other color. And it actually went through these efforts—you know there's the red, the taillights don't—you know there's not as much contrast, etc., on why red cars are less safe. And the person who wrote the article said people who buy red cars are self-selecting red cars, you know; it's as simple as that.

Moderator: Selection bias. The deign of the evaluators problem. Ruth, did you have anything you wanted to say before we open it to audience comments?

Ruth: Yeah. I just want to say, I think you know there's a kind of perpetual problem that we've encountered in this project with the word "evaluation," because it's really, it's such a broad word. It covers so many different kinds of knowledge generation. And what we tried in sort of a very disciplined way was to consistently talk about impact evaluation. But there's a whole range of other kinds of evaluations that are done within development institutions, within public sector agencies and NGOs, that have to do with extracting operational lessons about how projects worked, how—you know how well did the supervision work, how well did the government counterparts interact with—with agency staff, and a whole range of other sort of implementation-level questions that really can—really best be answered by looking—by having folks in the institution who really understand how it works do that work. So, I really don't want this to be misunderstood in any way as kind of a slam against the evaluators, or the evaluation departments in any major institutions at all. I personally think that people who work in evaluation in development are the heroes, because they're—you know it's not the glamour track to say the least. These are people who are really trying to squeeze knowledge out with very limited resources and with quite limited attention by management. And so, again, this is—the hope is that this can be a strong complement to that important work that goes on within the agencies and within developing country governments in many cases.

Moderator: Bill?

Bill: I'd just like to add two things. One is, I think Jon mentioned the benefits of sharing the technology of Impact Evaluation. This is something that specifically came up in, I remember

particularly in Mexico. Gloria Rubio, who is in the Social Development Ministry, was saying they're required by law to do impact evaluations of social projects and they're looking for ideas of how to do it and they don't know where to go. They have a few consultants they rely on a lot, but they would really benefit from this. And this came up in several—several of the consultations in the developing countries. And just one qualification that didn't quite come up on the panel, but a lot of people looked at this chart or this discussion and thought we were creating some kind of new evaluation research department. And I want to emphasize that the collective initiative here could be something as simple as a committee that's exchanging information; was on of the institutional options, or that a council of a handful of staff that are administering grants or administering review processes. This is in addition to what the agencies are doing and trying to leverage all that, but not to be some kind of new monopoly-central thing. So, I just wanted to make sure that was clear.

Moderator: Those are important points. Thank you. If you'd like to—in fact, there's nobody at the mic, so if somebody wants to go there. The others we'll take—okay, the three who are up and then the next, so you don't all have to stand at the mics. I'll take another round subsequently. Dennis?

Dennis: My name is Dennis de Tray. I work for CGD. Normally, I would not leap to it, but I think a very interesting issue has been put on the table that I would appreciate the panel's views on and it has to do with the conversation about the bank's evaluation department. Having been the recipient of a number of bank evaluations (not me directly), but as country director in my programs, let me tell you that I don't think the issue is independence. I think the issue is design. These evaluations are inherently incapable of producing counter-factuals because they're ex-post. So my—and the debate that takes place in the corridors may be about spin, but it's also about interpretation. These are enormously complex experiments with hundreds of reasons for success or failure. And what the evaluation department is supposed to be doing is deciding whether the bank's intervention worked. And you can imagine how hard that is even in a controlled design, let alone ex-post with no control design. So, my question to the panel and even to Ruth and Bill is, is the issue really independence, or is that we need to develop more rigorous, as Jon was putting, a system of evaluation? And by the way, having been at Rand in the '70s when that health experiment was being promoted, let me tell you, it's expensive in every dimension that you can think of. I really, really wonder if we realize how expensive.

Moderator: Next? We'll take a few. Please go ahead. We will take those who are currently up. We'll take four.

Speaker: I'm working for the Analysis Information Management and Communications Project. I think it's an excellent idea having an independent entity and I'd like the clarification made by David just now. Because I think one risk that one could encounter is that it's perceived as like evaluation police, and I think one has to be very careful to define such an independent entity. The questions related to it are really more about how would such an entity be linked to ongoing efforts to improve the quality of research and evaluation? This is a great effort, but schools of public health, international networks like INCLAN, have a similar goal. So, how do you complement, how do you value what's already going there. How do you link to these efforts and putting a seal of approval on some of the research that I think has passed a certain level of muster

and makes a lot of sense? I had a somewhat lighter question, and that was, if I would have gone around the panel and asked each of you how do you define "impact," I was wondering what you would have said?

Moderator: The lady in the back.

Jo Marie: My name is Jo Marie Griesgraber and I'm with the New Rules for Global Finance Coalition. And I think it's an interesting idea, and I think in terms of peer support, fostering competition of ideas, training; also a great library, an electronic library and how do you find every good possible evaluation that exists would be terrific. But my question is, how do you link up with work that's being done in the World Bank and a little tiny bit in the IMF on poverty and social impact assessment? There's a whole methodology that's been developed, an enormous handbook or bible. It's not referred to that I could find in your text. They do sectoral analyses and yours tends to look at education and health analyses is my sense of your approach. And the challenge—if you're looking at economic development, you do have to do—you have to be ex-ante in your projections, as well as looking at macro-economic policies. I mean I know health and education are enormously difficult, but to say that it's difficult excuses no one. So the need for ex-ante macro impact analyses, so countries can choose from policies, I wonder how you would fit that in your framework and to what extent you have been working with Band and Fund on their sectoral and system, or you know I guess sector-specific when they deal with energy or they do tax policy, to haven't done countrywide analyses yet. Thank you.

Moderator: And the gentleman behind you, Marie.

Paul: Thank you. I'm Paul Applegarth. I'm the Senior Transatlantic Fellow at the German Marshall Fund. But as some of you know, I do know a little bit about what MCC was trying to do in its design and origination. So if you'll let me indulge, there'll be a couple of minutes with a few comments and questions at the end.

First of all, I think you all should be applauded in your effort in even trying to undertake this. The more we can get a focus on impact on the ground and what's really making a difference in the lives of people and shifting discussion away from Congressional "which committee has jurisdiction" or "which agency is responsible for doing this or that" in this sort of inside the beltway issues, the better off we are. And the more we continue to capture the potato vine and the public dialogue about making a real difference in the countries, the better off we are and this clearly helps in that. In fact, I would share probably in Jon Baron's comments. These kinds of disciplines you would apply to domestic programs as well as international programs. A few of those could benefit from an evaluation that works.

There is a surprising lack of data of what works in our experience. That electronic library that you were asking for would be very small at this point. We found that early on when we were at MCC trying to figure out what did work and talked to the best people in the field. I had a great illusion that we'd have a composite early on of what works. It turned out it was—it turned out to pretty hard to do. And the more departments of countries would come to us and say, "Give us advice in structuring the impact; what works or not. What can you show us that really works or not?" The reality is that it's not there, and so again, this effort to try to design upfront into the

programs true rigorous evaluation models is very important and I would not lose that theme no matter how you work—what direction your work goes.

I think that part of the good news is that it really isn't very expensive. At least when you're talking about programs the size of MCC's or elsewhere, as a percentage of program costs to do a rigorous evaluation model with external parties is cheap; not in dollars perhaps, but in percentage of program terms. And yes, it takes time to get results, but development systems are around 50 years and we could certainly have a better body of results than we have now and would be far more beneficial to the developing world than we currently have.

An obstacle you didn't identify though is the pressure to disperse. Building and collecting that initial data, sort of the baseline data takes time, and with the pressure to get all the money out the door right away, it's a real obstacle. And so there's a need, I think, in the group for helping on education, getting those that have legitimate concern with that money that's been appropriated and hasn't been dispersed, to help them understand why that's true. And it isn't a case of your bureaucratic inertia or ineptitude, it's a question of trying to do it right and make sure the money is used fairly so that the program really achieves results. There is a key education need; not in the committees like Senator Lugar or Senator Hagel or Senator Biden or Jim Kolbe's committees, but the others out there. There's a real need to educate those that are concerned about how resources are being used and utilized well to make sure they understand that this program evaluation is key and it takes time to get the date.

I think that the most positive development and hope for the future is the involvement of the partner countries in this. Because they have a stake in the outcomes. They really want these programs to work and they are focused on results. They're not simply focused on commitment rates. For example—and I would use—and they are natural allies in trying to build this work going forward and I would use them whether it's Tim Thahane or there are several ambassadors that are in town, who will say exactly the same thing.

I'm surprised actually—normally I agree with David virtually everything, even though he was I guess our auditor at various stages. I don't agree with you on this though. I think you're overly cynical on the "no hypothesis" discussion at MCC. The compact outcomes are built in, desired target incomes are in the compact as soon as they sell, so they should be available. Whether or not they've been achieved should be available as a matter of public records, so it won't be a case of rewriting history; they've been achieved or not and it's an independent—and the public should have an ability to be able to evaluate whether the programs have succeeded or not. In addition and notwithstanding the overall programs, there are in at least some programs individual evaluation models really testing what works or not and I think the best one to look at is probably the Burkina Faso Threshold Program, which is clearly designed with quite serious modules of "do you train a teacher daycare" is the objective of the program is to increase girls' graduation rates; whether the best intervention is at the ground level to do that. And a variety of tests in the program itself to try to get serious arms' length data on that, so that the next time the program is done it will work better. So, I would be somewhat more sanguine about that than your comments imply.

Moderator: All right, are you coming to the-.

Paul: I am coming to the point. As I said, [unintelligible] indulgent—I'm sorry. I'm coming to—I guess the thing that disappoints me is a little bit the Manifesto and the attention given to the call for collective action and essentially into an institution, whatever form it is. Although, if I could be backed off a little bit, because both of those take time and they externalize the problem. There's a lot of agencies that have already built the cat that are trying to do things. It's just not MCC; [unintelligible] some things at the [unintelligible], some things at the Bank. There are examples where agencies are trying to do this and those who say it's too hard could benefit from having those examples made public and put the pressure on the existing agencies to transform themselves, rather than allowing those in this agency say, "Well, this external is being set up and I don't have to worry about it anymore." And I think I would like to see equal collective pressure on the existing agencies to reform their own operations to really build these in upfront and to design and comment. And I'm curious—this is my question—whether—how much consideration you gave to that and why you opted to go this direction rather than a heavy focus on the existing agencies and reforming their own operations? Thank you.

Moderator: Thanks very much. I said I'd take four, but you've been standing a long time, so if you promise to keep it brief, I'll invite your question.

Don: Thank you. I'm Don Shirk. I spent 15 years with the Treasury dealing with multilateral banks and for my sins I was on the boards of three of the regional development banks. I think it makes me about 30 times more guilty than you Ken.

Ken: Yeah, 30 times more knowledgeable, too.

Don: Let me say first of all that I appreciate the opportunity to hear this subject debated in depth and I think that Ruth and Bill should be congratulated for that. And I don't get this chance very often, but I want to tell people that I'm a member of the Paul Applegarth fan club for his work with the MCC and I'm sorry to see him step down.

The thing that's missing in my judgment is the subject of governance, and it can be clearly stated in two simple little examples. Why was it that the United States Treasury and government succeeded in getting the World Bank's evaluation department to report to the Board, but we were not able to get the Asian Development Bank to have their evaluation group report to the Board, as opposed to the President? And indeed in the case of the African Bank, they too reported to the President and not to the Board. It was specifically because this was an American issue and that the Americans—I've heard it 100 times in my career—Americans run the World Bank but not the other banks. The Japanese economist that some of you may know—I'm sorry, it's just gone out of my mind—he once told me the World Bank is America's bank and the Asian Bank is the Japanese "baby." It's the Japanese baby. In the case of the African Bank, the best example is as follows. We were—we required the management of the bank to bring evaluations to the Board for the Board to review and to criticize or to accept, whatever. There was an evaluation of an educational project for I believe the country of Senegal. It was to come to the Board and it was—we had a two-month schedule of Board events, so we knew when it was coming up and the evaluation for the Senegal—I'll go fast—the evaluation for the Senegal Education Project was

coming a month from now, but a new Senegalese project was up this week. I made the simply suggestion that since there was an evaluation of the African Bank's prior work in education in the same country, wouldn't it make sense to defer by only one month consideration of another educational bill—an educational project for Senegal? You would think that it was so obvious. I had to escape that Board room with my life. I was—I was called every name imaginable. I was anti-development. I had things against the poor people of the world, simply because I wanted to hold up and see what they said about the first evaluation.

So I think those two examples may suggest why I think governance has to be looked at. And for my way of thinking, the OECD might be a place that you would start to talk about an independent evaluation board. Thank you.

Moderator: Okay, clearly we have a range of knowledge and interest. I'm always impressed with the depth of knowledge in the audiences that we attract to CTD events. I'm going to summarize really quickly and then I think give Ruth and Bill the prerogative, but if any members of the panel would like to answer particular questions, you'll certainly have an opportunity to do that.

There was a question from Dennis de Tray which sort of, you know, isn't it a lot of it about design rather than independence. I didn't get the second gentleman's name, but it was concerned about linking to ongoing efforts to improve evaluation within other organizations. Jo Marie Griesgraber asked about—she endorsed the idea of peer support, having a library which somebody else said would be very small indeed if it was of rigorous evaluation. But her primary point was about the link to macro policies and the need for ex-ante planning about the impact of macro on social outcomes. Paul Applegarth, welcome. We're indeed delighted that you're here. I heard broad endorsement for the initiative in general, but some thoughts about whether or not the independence was necessarily the way to go and perhaps my colleagues on the panel caught the nature of that concern better than I did. Finally, Don Shirk was talking about education and governance, in particular; if it isn't the problem of governance with the multilateral banks. I've only touched briefly on elements of what our audience participants raised, but I leave it to you. Ruth and Bill do you want to hold off? Do you want to take any pieces of that? How do you want to go?

Bill: Splitting up the things here. First of all, thank you for those comments and as Lawrence said that the wealth of knowledge and expertise in the room is just incredible, so it's really helpful to hear the stories. And the entire consultation process has been like this. I've been astonished how much people are interested about the topic and then have incredible stories to tell. This is like a—sort of gives the rest of you a little bit of a sense of what I've been hearing for the last two years. I'm just going to respond to two the questions because there isn't time for everything.

First, Dennis' point about the issue of it being independence or design. I think you're absolutely right and that's why the working group, if anything was emphasized over and over again it was the quality of the studies. That there's a lot of money that goes into studies that don't end up giving impact evaluation knowledge. And when the design is really good, I think it speaks to some of the other points. When the design is very good, it constrains the amount of discussion

about interpretation as well and makes it harder to spin it too far in a direction or another. So, I really heartily support that. I think the working group really came to that conclusion as well.

And then the other point I was going to respond to was Paul Applegarth's question about whether it would be sufficient to do collective efforts to build pressure on existing institutions. And there is some possibility that that could work. We looked at that. Our concern was sort of in the group discussions were two aspects. One is, we're not convinced that the—that the current initiatives that we see starting and looking very good—the MCC—actually, I met David Amio on the plane on the way back from Cape Town and was very impressed with what they're doing, so it sounds very hopeful. And we also talked extensively with Gerlicher and [unintelligible] at the World Bank efforts, but we're not really convinced that they're sustainable. We've seen initiatives like this in the past and they don't seem to outlast particular visionary leaders or individuals. Maybe that's changing, but that was one concern. And then the second concern is that a lot of the organizations and countries don't have enough in-house expertise and capacity, or even the justification to get it to do this kind of work. And Patience Kuruneri, who was in our working group from the African Development Bank, said if something like this existed it would be great for them, because they could actually, in a sense, use the services of the council to contract the expertise to have the networks; something they can't justify having in-house. The World Bank is large enough. It has a critical mass of people to do this kind of work and I gather the MCC has funds for this. But when it comes to a small developing country government, a smaller bilateral agency or the regional development banks, it's not really clear they have the capacity and expertise in-house. So those were some of the reasons for that choice.

Ruth: I'll just take very briefly a couple of the other points. On—on independence, I think that there are many issues around this and I'll just sort of refer you back to think about that slide that Bill had talking about sort of splitting and linking and where the need is for ensuring that people who have a very vested interest in the program and really care about the answers have a strong influence over what questions are asked. But not necessarily on what—on how the finding are disseminated and shared. There is one feature of independence that I think is hugely important and that is—among others—and that is the credibility that's conferred by an independent evaluation. And this is a notion that is something that is very comfortable to people, certainly in the corporate sector and who think about corporate governance, that having third-party evaluations, third-party audits is you know sort of one—one version of that. Not the version we're talking about. But that those things—and David it referred it to you and the work at the GAO. That sort of arms' length relationship is something that really can confer credibility when—particularly when the stories are—or when the results are very favorable.

A couple of other comments; one on the cost issue. You know, we have to think about sort of the "compared to what" question, and so I think the relevant comparison is the cost of generating this knowledge compared to the cost in dollars wasted and lives lost or potential—potential not realized, but the cost of not having this sort of knowledge. So I think that's the relevant comparison and—and it should be one that its tradeoff should be assessed.

On the point of about looking at sort of country-level policies; tax policies, macro economic issues, maybe poverty reduction strategies and so forth. What we're really focused on here is, as I think was clear is, well, how effective are those dollars when they're—when they're really

spent to provide the services to implement the programs that touch people's lives. Because if you don't know how effective those programs are, then you know making sure that the government budget grows at some percentage every year because the overall economy grows at some percentage every year, you have to make some very, very strong assumptions that that's going to benefit people and lead to long-term sustainable and economic development. So what we're focused on is some ways is very microbe, but in some ways it's the very heart of the matter I think. And many of the larger scale policies are promoted with a kind of assumption that may well be false, that—that the positive impacts on people will—will sort of automatically be generated if there are more government resources in particularly the social sectors.

And then one final small comment on how this would be linked to existing activities. I think the way this sort of initiative is or could be designed, is to be very—have part of it very much a sort of clearing house and have a networking function among many things that are going on. And you know just referring to the OEC DAQ evaluation network, which is a valuable, or an entity and serves a function of linking the evaluation departments, particularly in bilateral agencies. What it misses is, for example, the research departments in multilateral institutions where most of the impact evaluations are actually being done; misses the evaluation work that's being done in developing countries; misses the NGOs; misses things like INCLAN; and so you know, once possibility is to have that DAQ structure expand greatly. But OEC DAQ as an institution isn't, I think, quite organized to do that. An alternative is to have the sort of other entity that is explicitly intending to link these different functions. So, sort of hopefully a response to most of the questions that were asked.

Moderator: Is there any burning desire to reply from the panel? Yes.

Speaker: I wanted just to take a moment to respond to one gentleman and his question said we're not—we're not looking for an evaluation police and there was a kind of general nodding of heads "that's right." And then he went on to say what we're looking for is a seal of approval, and then there was a lot of nodding of heads. And—and you know, when I hear these kinds of discussions, often the same people who are talking about more local ownership, more control, often in contrast to our colleagues at the World Bank or the IMF—and you hear boos and hisses—you know those are the ones who are telling people what to do. And what we need to do is have more local ownership and more local control and they will applaud this. And I think there should be no mistake that what we're talking about here is a serious upgrading in technical—in the technical skills brought to bear on important issues that can't be undermined by ideological attentions to the things that matter, but that may undermine the value of the results. You know, Esther Duflo was part of this process. I don't know Esther. I don't even know if she's here. Even within the United States there probably aren't a large number of people with the technical skills that Esther Duflo has. And I think we got her from France didn't we? You know you go to do this in Asia or Africa and most countries may not have single person who can do that. And so, if you're talking about building capacity and working with those people, I think that that's absolutely—that's absolutely fine. But if what you're talking about—and again, going to why investing this money matter—it matters because the answers matter and not the process. Now, the process may—there are better and worse ways of doing the process. But again, what we're talking about is exactly a research police, not with guns but with seals of approval and with presumably the negative as well, withholding the seal of approval from research that

doesn't—that doesn't cut mustard. The frustration comes when you look at the UN report that looked at its own research and found 95 out of 97 research reports didn't—you know weren't useful. I mean we aren't talking about technical enforcement by intechnicrats who are very good at what they do and insisting that the money that's being allocated to this is done according to international best practice. And if we all agree to that, that's fine, but we shouldn't mislead ourselves in terms of thinking that these things aren't running, in some sense, at cross purposes.

Moderator: Ruth, do you want to respond to that? I can see—I can see the wheels turning.

Ruth: Well, you know, I was just thinking back over the many, many, many conversations we've had about this project and what it would mean to have standards of evidence, what it would mean to have some kind of external support to do impact evaluations in various countries. You know we can only talk to the smallest, the most non-random sample of people, you know of people who come to meetings that are called impact evaluations; not a random sample. And the resistance to the high technical—the idea of high technical standards, the resistance to having people who have expertise in particular methodologies, really has been concentrated in many of the development agencies. And I would say not at all, or very little—I mean don't recall any—among the people who have commented from developing countries. In sharp contrast, what we heard over and over again was the—the value that people who work in evaluation units in public agencies and developing countries and in NGOs and even at a commercial bank that does social investment projects; what we heard over and over again was a kind of "hunger" for having collaboration from outside experts who can bring lessons from other countries, who can bring perhaps you know the sort of most up-to-date methods. A hunger for that to do it in collaboration, but to do it as well as possible. And what we also heard was a tremendous—and I'm going to use the word again—a tremendous frustration with the consulting business in general and with evaluators who come for two weeks and who are contracted by development agencies who helicopter in and helicopter out, write the report, nobody sees it again and nobody learns anything from it. So you know, I'm making a stylized point here, but I don't think that there's necessarily a conflict between responding to local needs and demands and having high technical quality. And I'm not I'm capturing what you were arguing about, so anyway but.

Moderator: I think we need to speed-.

Ruth: Okay, I'm sure we wouldn't disagree.

Speaker: I know this could go on forever, but there's actually just one point I would—if I could follow up on.

Moderator: I think I'm going to end it. This is going to be an extraordinary event in DuPont Circle think tank realms and it's going to end on time. There are a couple of you who I had signaled that I would take your question. I apologize. This is clearly not a discussion that is going to be solved here today, but I would encourage you to stay on. I think we still have tea and coffee and cookies in the back and to exchange views with the panelists and our speakers. I'd like to thank the panel for joining us today and Ruth and Bill for your presentations. And thank you all for coming.