ReTHINKING EVALUATION

Reflections from Caroline Heider
Senior Vice President and Director General
Independent Evaluation Group
World Bank Group

#WhatWorks

Relevance

Sustainability

Efficiency

Effectiveness

Design quality

Impact

WHAT WORKS

IEG Independent Evaluation Group | WORLD BANK GROUP
World Bank • IFC • MIGA

1818 H Street NW
Washington, DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

Foreword

In a recent blog series, Caroline Heider, Director General Evaluation at the World Bank Group shared her thoughts on the time is ripe for the evaluation community to revisit the evaluation criteria that most development organizations use today. Over 100 of the series' 12,000-plus readers shared their comments and questions.

In this handout, we reproduce the blogs, along with a follow-up conversation with Ms. Heider and Hans Lundgren, Manager of the OECD/DAC Network on Development Evaluation.

Table of Contents

Foreword 1
Rethinking Evaluation – Have we had enough of R/E/E/I/S? 3
Rethinking Evaluation – Is Relevance Still Relevant? 5
Rethinking Evaluation – Agility and Responsiveness are Key to Success? 7
Rethinking Evaluation – Efficiency, Efficiency, Efficiency 9
Rethinking Evaluation – What is Wrong with Development Effectiveness? 11
Rethinking Evaluation – Assessing Design Quality 13
Rethinking Evaluation – Impact: The Reason to Exist 15
Rethinking Evaluation – Sustaining a Focus on Sustainability 17
Conversations: The Future of Development Evaluation 19
Biography 24

"Have we reached a Copernican moment where we realize the 'earth isn't flat,' and our definitions and 'understanding of the world' need to be reset? Leaving aside jargon and methodological challenges, there are other good reasons to revisit the evaluation criteria we use."

JANUARY 10, 2017

Rethinking Evaluation – Have we had enough of R/E/E/I/S?

After nearly 15 years of adhering to the DAC evaluation criteria, is it time for a rethink?

Over the past 30 years, evaluation in the development field has gone through multiple cycles of questioning which method is better than another. But few in the development circles in which I have operated, have questioned the standard evaluation criteria that we use.

Many development institutions, including the World Bank, regional development banks, the UN, and bilateral aid agencies, subscribe to what has come to be known as the DAC evaluation criteria. Specifically, these are five criteria – relevance, effectiveness, efficiency, impact, and sustainability (R/E/E/I/S) – that underpin most evaluation systems in international development.

Evaluation questions get framed around these criteria, and reports get written up using this language. But, many an evaluation struggles to implement these criteria in sincerity. Others are accused of using too much jargon as they report faithfully on these criteria. And often, the evaluations tend to leave readers with unanswered questions.

After nearly 15 years of adhering to the DAC evaluation criteria, is it time for a rethink? Have we reached a Copernican moment where we realize the "earth isn't flat," and our definitions and "understanding of the world" need to be reset? Leaving aside jargon and methodological challenges, there are other good reasons to revisit the evaluation criteria we use.

Values

As our societies develop, norms and values shift. Although the evaluation criteria appear to be neutral and should be applied as such, they were informed by a set of values. The post-2015 agenda has declared its intention to be more inclusive, respecting underprivileged groups of people, which means we as evaluators need to reflect whether the criteria these intentions. Being able to shape norms that are more inclusive of diversity rather than judge everyone through more limiting norms will be a necessity if 2030 is to become the world we want.

End Game

The adoption of the Sustainable Development Goals (SDGs) signals that we need to shift our understanding of development outcomes. Our development and economic models are premised on ever-increasing consumption. By contrast, the SDGs recognize that such consumption levels are unsustainable from an environmental, economic, and social point of view. This new commitment should lead to a paradigm shift around desirable development pathways that are not premised on escalating consumption patterns. Evaluation tools to unpack intrinsic impacts on consumption patterns will be needed to determine whether the world is evolving in desired ways.

Complexity

The world has become more complex, or rather: our ability to accept and understand complexity

has increased. International development has relied on often linear and simplified logical frameworks or results chains that string inputs-activities-outputs-outcomes-impacts into a straight causal path. Development practitioners, as much as evaluators, know that development processes do not follow such linear assumptions. Instead, one action might cause a number of reactions that have effects in rather diverse ways. Hence, we need to develop evaluation models that capture the effects of complexity to inform policymakers and practitioners about the actual effects of choices they make and actions they take (see excellent book on this topic by Jos Vaessen et al).

Technology

The pace at which technology develops and influences lives has far-reaching effects on societies. Solutions to complex problems can be generated in unthought of ways and often through unconventional networks of people. Information travels, is demanded, and influences large groups of people at a much faster and inter-connected pace than ever before. We are faced with an avalanche of data, a dearth of facts, and an ease of spreading (mis)information that has been unprecedented. Evaluation can benefit from technology, be it to construct with greater ease models that reflect theories of change, help with data collection and processing, or sharing evaluation evidence with a much wider audience than before. But it does so in an environment of multitudes of realities that may or may not lead to evidence-based decision making, especially if a "post-fact" era were inevitable.

Cost & Benefits

Current considerations of efficiency, cost savings, or cost-benefit analyses are challenged to take long-term impacts into account. Something that appears efficient today, might have inadvertent devastating long-effects on natural resources or the social capital of communities. Likewise, the distribution of cost and benefits have been uneven, as witnessed by those who bear the brunt of eroded natural resources, or of development outcomes that benefit some groups in society and not others.

Do these issues really necessitate a Copernican shift in the evaluation field that would require questioning the established five evaluation criteria? Are the criteria so inflexible that they can't be adapted as they are to address these challenges? Does this even matter for anyone else, other than the nerdy evaluators and their jargon-filled reports?

I say yes to all three questions. And particularly so, in a world that lives by the mantra "what gets measured, gets done."■

"As evaluators we need to shed light on whether an intervention's focus is on nodes in the network that matter, that can have large multiplier effects, or that are peripheral to the desired solution. That is a lot more than 'relevance.'"

JANUARY 17, 2017

Rethinking Evaluation – Is Relevance Still Relevant?

Meeting the bar for relevance is not all that hard, so should it be replaced with something more suited to a complex development environment?

In last week's #WhatWorks post, I argued that it was perhaps time for us in the evaluation community to rethink our evaluation criteria. After nearly 15 years of applying relevance, effectiveness, efficiency, impact, and sustainability as our foundational evaluation criteria, is now the time to change or adapt?

The evaluation criterion "relevance" has troubled me for quite some time. In many development settings, a project is considered relevant when "the aid activity is suited to the priorities and policies of the target group, recipient, and donor." Of course, this is important. In plain language, it makes us question whether the intervention aimed to address real needs.

But that is exactly where the challenge lies: the needs of whom?

In an ideal world, the needs of the target population are aligned within the community, with the government's priorities, and the policies of donors. In reality, such a theory makes a large number of assumptions. For instance, that the target community is homogenous, which it often is not. Nor are priorities at central and decentralized levels identical, whether for a real difference in needs or for political reasons.

In practice, evaluators often use policies of governments, donors, and aid agencies to assess whether an intervention is relevant in that context. More often than not, these policies are written in ways that can justify a whole slew of different activities. Hence, meeting the bar for relevance is not all that hard.

In addition, I would argue, this criterion might be irrelevant in today's world of complexity.

Look at network analyses that map out situational problems and how they are interlinked. The TED Talk by Eric Berlow illustrates in less than four minutes how complexity theory and technology allow us to map and understand development challenges in completely new ways. Being a visual person, I am fascinated by the modeling capacity that technology now provides.

More importantly, techniques like these could change the process through which we seek and find solutions to development challenges. They provide us with an opportunity to live up to the values of a more inclusive world, where the voices and perspectives of a much broader group of people matter in defining goals, solutions, and pathways that will get us there. This modeling capacity could help bring together the views of a broader set of stakeholders, add perspectives to

understanding a particular development challenge and interrelated factors, and come up with different solutions than, say, the solutions a group of experts might see from the vantage point of their technical expertise.

Moreover, an approach like this can help anticipate potential amplifiers of success, or what we used to call "killer assumptions" that are strong predictors of failure or diminished development outcomes. These assumptions are often embedded, unrecognized, in project or policy design.

Impractical? Watch the video and look at the model the US military had developed for the situation in Afghanistan. Berlow maps all of these factors into an interactive model and then identifies nodes that have much larger ripple effects throughout the system than others.

What does all of this have to do with the simple evaluation criterion called relevance?

If we apply relevance to a more complex reality in the same way we have used up to now, with the policy context as the yardstick to assess relevance, any intervention will meet the criterion as long as it falls anywhere in the network of interrelated factors.

But that is not important for decision-makers! Instead, as evaluators we need to shed light on whether an intervention's focus is on nodes in the network that matter, that can have large multiplier effects, or that are peripheral to the desired solution. That is a lot more than "relevance."

Instead, I suggest that we fundamentally rethink the "relevance" criterion and replace it with something that helps assess whether:

- Diverse perspectives were taken into account in identifying and implementing solutions, namely the networked analysis of the development challenge captures parameters that are outside a linear project logic that are essential for success or failure of the intervention;

- Development interventions address key entry points – the significant nodes that are bottlenecks or opportunities for multiplier effects – in a networked analysis of the development challenge at hand; and

- There are synergies across – or joining up of – a multitude of interventions aimed at the same development challenge.

Doable? Add your thoughts on what it would take. ■

"The pantheon of evaluation criteria – relevance, effectiveness, efficiency, impact, and sustainability – does not address the question of whether timely and responsive course-corrections were made when needed. In today's world – with a "new normal" of rapidly changing contexts, be it due to political economy, instability and involuntary migration, or climate change – this might seem surprising. But, 15 years ago development contexts seemed more stable, and the pace at which they changed was (or appeared to be) much slower than today."

FEBRUARY 7, 2017

Rethinking Evaluation – Agility and Responsiveness are Key to Success?

In many situations, stakeholders would benefit greatly from evaluative evidence that answers questions about the timeliness and appropriateness of course corrections.

Regular readers will recognize this piece as part of a series of blogs that discuss the challenges and changes that evaluation needs to live up to in the near future if it wants to avoid becoming redundant. For those who are joining the series now, please have a look back at our first two Rethinking Evaluation posts - Have we had enough of R/E/E/I/S?, and Is Relevance Still Relevant? - and join the debate by commenting below. We are looking for your ideas, feedback, and debate.

Development practitioners have, for some time, argued that they are held accountable to objectives set several years earlier in a context that might have changed dramatically since. We evaluators, in turn, suggest at least two arguments in return. The problem might arise from poorly defined objectives at the outset that did not allow the flexibility to adjust tactics during the pursuit of a higher (and still valid) objective. Or, in the absence of redefined objectives, it is not clear when or what kind of course corrections were actually introduced that would provide the new basis for evaluation. Rigid bureaucratic systems often create disincentives to revising objectives, or misunderstandings exist about how changes to objectives are reflected in evaluations.

Even if we resolved these problems, however, the pantheon of evaluation criteria – relevance, effectiveness, efficiency, impact, and sustainability – does not address the question of whether timely and responsive course-corrections were made when needed. In today's world – with a "new normal" of rapidly changing contexts, be it due to political economy, instability and involuntary migration, or climate change – this might seem surprising. But 15 years ago development contexts seemed more stable, and the pace at which they changed was (or appeared to be) much slower than today. Hence, the leaders in evaluation did not think, at the time, about the need for assessing agility and responsiveness.

This gap has been a larger issue in the humanitarian world. Rapidly evolving emergency situations need timely responses and challenge responders to be agile and responsive to constantly changing situations. In these situations, stakeholders – from managers who must make

quick decisions to donors who need to prioritize scarce resources – would benefit greatly from evaluative evidence that answers questions about the timeliness and appropriateness of course corrections.

This area, however, is a poorly recognized and hence hardly satisfied demand. Evaluators could address it by adapting questions and tools of the craft. Questions that could enter the evaluator's repertoire could include:

- Was the need for change anticipated at project design? Clearly, this is not the case for sudden-onset disasters like earthquakes. But in other cases, an evaluation should be able to determine whether the potential need for changes in the future were recognized and built into adaptive management and corresponding monitoring systems.

- What drove the adaptation process? Here, an evaluation should seek to understand whether development partners proactively monitored relevant indicators and situational information and how that information was used in deciding on course-corrections.

- Was adaptation timely? Establishing timelines of events and tracing when course corrections were undertaken will be essential to determine whether solutions were sought proactively or rather forced by circumstances.

- And what would have happened if….? This is a classic question of establishing counterfactuals, but in this case is needed to determine whether outcomes were better or worse because course corrections were made or failed to be made.

These are tough challenges to grapple with in evaluation, particularly because many of the details, processes, and conversations that lead to course corrections are not documented.

Nonetheless, because agility and responsiveness are important determinants of success or failure, evaluation needs to adopt a specific focus on agility and responsiveness to provide feedback, by giving credit for responsiveness and agility when it is due, and, when needed, identifying opportunities to improve. This alone, I believe, will incentivize debates and actions within institutions to anticipate the need for timely and responsive adaptation.

Will that be enough to overcome inertia where it exists? Maybe not, but it is a contribution that evaluation can make. ■

"One could think that evaluating efficiency does not matter, in spite of resource scarcity and the ever increasing need for improved cost-effectiveness. However, if anything we need to get better at assessing efficiency for a number of reasons."

FEBRUARY 28, 2017

Rethinking Evaluation – Efficiency, Efficiency, Efficiency

In times of resource constraints – have there ever been days without? – one would think "efficiency" would be at the top of the agenda for almost everyone. Unfortunately, we have seen limitations to this evaluation criterion in definition and above all in practice.

Efficiency is often defined in terms of "measuring the outputs – qualitative and quantitative – in relation to the inputs. It is an economic term which signifies that the aid uses the least costly resources possible in order to achieve the desired results. This generally requires comparing alternative approaches to achieving the same outputs, to see whether the most efficient process has been adopted." (OECD/DAC key terms for evaluation)

Way back when I was evaluating development projects at the Asian Development Bank, we used a definition that focused on the economic efficiency of projects; a practice shared across multilateral development banks. It is implicit in the above definition (note the reference to the economic term and least-cost models). It is calculated as economic rate of return, and uses a "net present value" of the investment – a standardized rate – to determine efficiency against alternative investment opportunities. This approach goes beyond the narrow definition of efficiency that compares input-output relationships, maybe more often used in grant-funded aid projects.

But, as pointed out in an IEG evaluation of 2010, the practice of Cost-Benefit Analysis has been on the decline at the World Bank for several decades, dropping from 70 percent of projects including calculations of economic rates of return in the 1970s to 25 percent in the 1990s. This drop was in part explained by an increasing number of projects in sectors for which this kind of cost-benefit analysis was not feasible. Even when undertaken, the results of the analyses were not used in deciding whether to fund a project or not, undermining the rationale for undertaking the calculations in the first place. Another study, commissioned by the German Ministry of Development Cooperation, compared methods to assess efficiency used both at appraisal and evaluation. It concluded that many methods were little known and used.

One could think that evaluating efficiency does not matter, in spite of resource scarcity and the ever-increasing need for improved cost-effectiveness. However, if anything, we need to get better at assessing efficiency, for a number of reasons.

The systems approach that complexity requires us to use has the potential for comparing different intervention options and a combination of them. Let's assume we could model a development challenge just like the US Army had with the conflict in Afghanistan (see TED Talk by Eric Berlow); it could allow development practitioners to identify not only options that would generate

the highest impact, but also options that are more or less costly, and to determine the most cost-effective package of interventions. Evaluation could assess the quality of those assessments, and whether they were used in decision making, and could complement the estimates made at design with data on actual costs and benefits at the time of evaluation.

More immediately, there is a great need to factor into the cost of interventions the hidden costs of social and environmental impacts. Today, the cost of pollution is more often factored into investments, especially when mitigating measures have to be taken or technology has to be adapted to clean up pollutants rather than releasing them unfiltered into the atmosphere. But more will need to be done in evaluating the efficiency of these investments over alternative choices.

Finally, evaluation methods for efficiency will need to become more sophisticated to deal with waste. Losses, such as in electricity or water distribution systems, do get accounted for in the evaluation of economic efficiency. However, as the SDGs call for a change in consumption patterns, methods will need to develop a better understanding of the consumption patterns implicitly (and hopefully increasingly explicitly) that an intervention promotes, determine when they are wasteful, to signal the need for rethinking of incentives.

Is evaluation ready to rise to these challenges? Comment below and share your opinion with us. ■

"If development planners were to use complexity models to understand the web of interrelated processes to identify their objectives, intended and possible unintended effects would become clearer, and possibly increase evaluability."

MARCH 21, 2017

Rethinking Evaluation – What is Wrong with Development Effectiveness?

The way we look at development effectiveness needs a facelift.

Effectiveness is central to international development and its evaluation. The OECD/DAC Glossary of Terms defines development effectiveness as "the extent to which a given development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance."

By itself, the DAC definition embodies the accountability dimension of evaluation. Complemented with an evaluative question of "why" objectives were achieved (or not), one gets to learning about the experience of trying to achieve a particular objective in a particular context.

The term embodies the fundamental concept that development assistance is measured against the yardstick that it sets for itself, because it is the development partners who decide on the objectives they aim to pursue. This notion is very different from assessment tools like benchmarking (a comparison with an agreed standard) or competition, where success is defined in comparison with others.

When viewed among these options, effectiveness seems rather lenient, given that the development partners define what success looks like. Nonetheless some development practitioners argue that effectiveness is too tough, and too rigid to account for adaptation during the life of the intervention (read our earlier blog – Rethinking Evaluation: Agility and Responsiveness are key to success). Others, mostly evaluators, argue that it is the practitioners' risk aversion that makes them shy away from effectiveness as a measure of accountability, and has incentivized behaviors to "game the system." In that scenario, objectives are written to get a good rating at the end rather than as the intended results that development partners try to achieve. There are good points to each of these arguments.

But, from my perspective, there are additional reasons why the way we look at development effectiveness needs a facelift!

With our increasing understanding of and ability to work with complexity there will be different demands on project planners and evaluators, as discussed in an earlier blog about relevance. This might change the way in which objectives are set, which will either make it more challenging to assess whether they were met, or demand an equally dynamic evaluation tool, or both. It raises questions about the differentiation between effectiveness and impacts – something many practitioners have struggled with – and might call for merging these two criteria.

In addition, the way effectiveness has been defined has kept attention focused on intended

results. Most evaluations grapple with getting evidence to determine whether objectives were achieved and to measure an intervention's contributions. Fewer evaluations are able to collect evidence on effects outside the immediate results chain and identify unintended consequences. If development planners were to use complexity models to understand the web of interrelated processes to identify their objectives, intended and possible unintended effects would become clearer, and possibly increase evaluability. And even if planners do not use such tools, evaluators should explore how they can become part of defining program theory and evidence collection.

At the same time, complexity models make it clearer that attributing change to a single actor or intervention ignores that many forces are at play. The question of attribution has been at the heart of many a debate about the rigor and validity of evidence and whether it could be proven that one policy or action was better than another. A better understanding of complexity might help join up interventions of different development partners, and suggests that (in the long term) evaluations have to be undertaken from a systemic point of view rather than focused on a single development agency or intervention.

Likewise, distributional effects of interventions, whether explicitly part of the intended outcomes or not, need to be assessed if we are serious about goals like "no-one left behind" (proclaimed by the global community through the SDGs), or boosting shared prosperity, as one of the goals of the World Bank Group. Too little attention is paid to the assumptions we make about interventions that are not targeted and supposedly have no distribution effects. If the analysis of intended and unintended effects is differentiated by different stakeholder groups (rather than "beneficiaries" as one homogenous category), we can get a better understanding of the actual effects or impacts of interventions.

In short, the criterion "effectiveness" needs a facelift, not just for the purpose of addressing counterproductive behaviors. The spotlight that we evaluators shine has incentivized certain behaviors of decision makers, program planners and implementers. Let's do so intentionally, rethinking evaluation criteria and methods that incentivize behaviors for better development outcomes. ■

"An assessment of the design quality of an intervention asks whether the 'intentions were right.'...But, will this be enough as we prepare for the future?"

APRIL 11, 2017

Rethinking Evaluation – Assessing Design Quality

Will asking if an intervention's "intentions were right" be enough as we look toward the future?

Evaluators in international development typically assess the design quality of an intervention. The reason is simple: the intervention design provides the yardstick to determine success and failure. Achievements are measured by comparing what was planned with what has actually been achieved. In addition, we find that "quality at entry" (in other words: the quality of intervention design) is a good predictor of the intervention's outcomes.

An assessment of the design quality of an intervention asks whether the "intentions were right." Questions we evaluators use to determine the answer include whether:

1. Objectives were realistic,
2. There is an internal logic or coherence along the results chain that would ensure inputs and outputs could actually lead to the expected higher-level results, and
3. Relevant measurable indicators were embedded in design and monitoring systems.

But will this be enough as we prepare for the future?

Previous blogs in this Rethinking Evaluation series discussed issues that need to be reflected upon in both intervention design and its evaluation.

For example, in an earlier blog on effectiveness, we argued that a different approach to managing complexity might lead to a different way to define objectives; one where it matters to understand the web of interrelated factors so as to identify entry points that will amplify possible impacts and be cognizant of what have been, until now, unintended consequences. Under these circumstances, is it enough to ask whether objectives are realistic?

Or take the Relevance of Relevance blog where I argue that the simple question as to whether something is relevant in a complex network raises additional questions about approaches that have depended on a rather linear interpretation of reality. Logical models and results chains, if actually used in all sincerity, have far more often than not been simplified and linear. Question 2 above is clearly aligned with that question of internal logic.

There are additional challenges that intervention design needs to take into account in light of the Sustainable Development Goals (SDGs).

For instance, tensions can and will arise, especially between goals that require tough trade-offs; a challenge that is embedded in the SDGs that want a better life for all, but in environmentally sustainable ways. In an ideal scenario, the tensions between goals will stimulate innovation and lead to better solutions. Say, lowering of costs of alternative sources of energy to ensure we can meet goals to give equal access to electricity to all without, however, further depleting the earth's natural resources. But these ideals might be hard to attain. Evaluating whether and how trade-offs were weighed and whether they influenced ultimate intervention design will add a much better understanding of whether the right decisions were made.

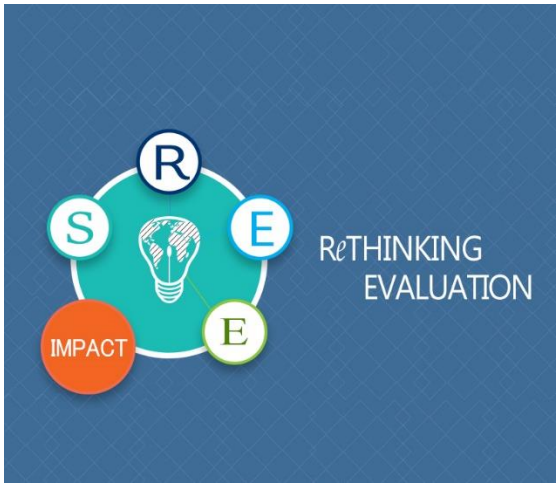
And then there is the need to evaluate whether

- Diverse perspectives were taken into account to identify and build into the intervention design a focus on central levers of change. In addition, understanding whose perspectives were taken into account is important to understanding ownership and how stakeholder groups will be affected.

- Features were included in the intervention design to track and respond in a timely way to changing contexts, as discussed in Agility and Responsiveness, be they to manage complex (or complicated) political economies, or operate in dynamic institutional contexts.

- Interventions have intended or unintended, direct or indirect effects on consumption levels and patterns as suggested in SDG12 and discussed in the efficiency blog. Introducing measures to evaluate this dimension now will create greater awareness and incentives to change project designs and implementation.

These factors (and more) need to be reflected in design quality and its assessment, whether through internal quality assurance processes and evaluation. If we do not start now, necessary evidence will not be generated in time to learn from experience and make course-corrections as they are needed. ■



"In spite of considerable resources spent, the quality of too many [impact evaluations] is not high, the results deemed not conclusive and limited to rather narrow phenomena, while leaving fundamental gaps on strategic issues. More often than not, these studies conclude that more studies are needed. Confronting this reality, as well as evidence about the weaknesses in project design – poorly defined objectives, confusion between outputs, outcomes, and impacts, and ineffective M&E Systems – together with insights into complexity theory, gives me pause to think!"

MAY 2, 2017

Rethinking Evaluation – Impact: The Reason to Exist

Complexity theory and enhanced modeling capacities provide opportunities to rethink evaluation methods.

Long-term evaluators in the development field will remember the difficult conversations we have had (not too long ago) about measuring impact in a reliable way. The reason for heated debates is simple: positive impact is what development interventions are meant to produce, and negative impact is what they are supposed to avoid—and proving it one way or another is paramount.

Impact is defined as follows: "the positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The examination should be concerned with both intended and unintended results and must also include the positive and negative impact of external factors, such as changes in terms of trade and financial conditions." (OECD/DAC key terms for evaluation)

Methods for impact evaluation have grown over the past decade. A whole industry has sprung up with many a student leaving university with great aspirations to undertake impact evaluations of a certain kind. But, as many systematic reviews and a 2012 IEG evaluation of the impact evaluations undertaken by the World Bank Group show us: in spite of considerable resources spent, the quality of too many of these studies is not high, and the results are deemed not conclusive and limited to rather narrow phenomena, while leaving fundamental gaps on strategic issues. More often than not, these studies conclude that more studies are needed.

Confronting this reality, as well as evidence about the weaknesses in project design – poorly defined objectives, confusion between outputs, outcomes, and impacts, and ineffective M&E Systems – together with insights into complexity theory, gives me pause to think!

Let's assume development practitioners take the opportunity that complexity theory and enhanced modeling capacities provide – something that I believe will have to happen. Let's also assume that such a change will result in getting to a better understanding of development challenges, pathways to their solutions, and interventions that are designed in different ways (as argued in my

earlier blog What's Wrong with Development Effectiveness?) – ways that recognize the systemic effects interventions can have on a more complex network of interrelated development processes. It is hard to imagine how a logical framework or a traditional M&E system would capture impacts as defined in the DAC evaluation criteria, let alone the cost of doing so.

Instead, we evaluators need to seize the opportunity to rethink our practice. Evaluation methods and questions can continue to incentivize changes in development practice. This could be effected by:

- Showcasing how complexity models can be used in evaluation and, hence, applied to design;

- Asking evaluation questions that move beyond linear results chains into areas of unintended direct and indirect effects that interventions may have; and

- Strengthening methods to capture synergies between interventions, and taking a systemic perspective of sets of development interventions.

Some thinking has gone into what complexity means for evaluation practice. One excellent reference, for example, is Dealing With Complexity in Development Evaluation, authored by Michael Bamberger, Jos Vaessen and Estelle Raimondo. But a lot will need to be done to translate these ideas into evaluation practice.

"Taken together these dimensions of sustainability – economic, fiscal, environmental, and social – are complex. It will be hard and costly to try to address them systematically in all evaluations. At the same time, we evaluators cannot afford to turn up with empty hands and concerns about missing data."

MAY 30, 2017

Rethinking Evaluation – Sustaining a Focus on Sustainability

Looking back on years of using the sustainability evaluation criterion, one has to ask – how well have we done?

The UN Sustainable Development Goals (SDGs) have brought renewed attention to sustainability. Although the DAC evaluation framework includes sustainability as one of its five criteria, looking back on years of using this DAC evaluation criterion, one has to ask – how well have we done? And here I mean in evaluation practice rather than results.

More often than not have I seen sustainability used in different ways than it was originally conceived. The definition - "[s]ustainability is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn. Projects need to be environmentally as well as financially sustainable" (OECD/DAC key terms for evaluation) - focuses clearly on the outcomes of the intervention and their sustainability.

Many evaluations, however, assess whether the projects themselves will be sustained, often concluding this to be the case when funding is secured from government or another donor. That is right, especially for facilities that continue to be run by the public sector and require government funding. But, the same is true for public-private partnerships (PPPs) where a recent evaluation of ours showed that the impact on government expenditure (in other words: fiscal sustainability) was hardly ever assessed.

Sustainability is also often taken as synonymous with environmental sustainability. When I was leading project evaluations, we hardly ever had the time and resources to assess environmental impacts and whether a project would leave a lasting footprint, positive or negative. At IEG we are in the process of evaluating environmental pollution projects in the World Bank Group, which will shed some light on past practices, including data that is available today and remaining gaps.

Under the SDGs environmental sustainability goes much further than a "simple" question of pollution. It is about the use and depletion of natural resources, about consumption patterns that are out of bounds, and the distribution of consumption patterns. For instance, when we look at access to electricity, our recent evaluation showed how underserved countries, especially in Africa, are. Sustainable Development Goal 7 is committed to expanding access to affordable and clean energy, increasing renewable energy sources, and attaining energy efficiency as a measure of improved consumption patterns. But it will not be sufficient for evaluations of the

power sector to assess efficiency gains that must be achieved in other parts of the economy.

Closely linked to the World Bank Group's goals of poverty reduction and greater shared prosperity is the question of social sustainability. Upheavals during the past years have often been rooted in growth that has excluded a broader base, where wealth, access, and voice have been captured by the few. The commitment to inclusive growth necessitates that we understand better the distributional effects of interventions, whether they were designed to target groups previously excluded or not. Almost more important for us is to evaluate and understand interventions that we believe to have no distribution effects, to shed more light on the actual distribution of results that they have. IEG is in the process of evaluating the World Bank Group's experience in this area to generate some early insights.

Taken together these dimensions of sustainability – economic, fiscal, environmental, and social – are complex. It will be hard and costly to try to address them systematically in all evaluations. At the same time, we evaluators cannot afford to turn up with empty hands and concerns about missing data. We need to debate how we would evaluate interventions through these lenses of sustainability, see that the right questions are asked during the design of interventions, and incentivize the collection of relevant data. ■

JUNE 21, 2017

Conversations: The Future of Development Evaluation

A candid conversation as to whether it is time to re-think DAC evaluation criteria – relevance, effectiveness, efficiency, impact, and sustainability; in short R/E/E/I/S – that underpin most evaluations in international development.

The growing interest in strengthening development outcomes has stirred increasing debate about evaluation effectiveness. Today, many development institutions subscribe to what has come to be known as the DAC evaluation criteria. Specifically, these are five criteria – relevance, effectiveness, efficiency, impact, and sustainability; in short R/E/E/I/S – that underpin most evaluations in international development.

As a follow up to the blog series, Ms. Heider and Hans Lundgren, Manager of the OECD/DAC Network on Development Evaluation, took part in a conversation where they shared their thoughts on the state of development evaluation today.

Question 1: *Let's start with you Caroline. For many years, the development community has used a common set of evaluation criteria, commonly known as the DAC evaluation criteria. In one of your recent blogs, you suggested that now is a good time to revisit the DAC evaluation criteria, and that we may be at a "Copernican" moment. Why do you think so?*

Caroline Heider: Copernicus is a famous symbol for rethinking how we see the world. For a long time, models have been developed that made assumptions or simplifications. These assumptions were necessary to make the models work, but removed them from the complexity of reality. Today, we are increasingly able to cope with complexity, at least in our thinking and in our modelling capacity. Therefore, it is (in my view) time to move to development models – theories of change – that are less linear, more representative of complex realities, and build on adaptive management. These approaches require evaluation to become more dynamic as well, adopt methods that capture complexity and unintended effects. In addition, there is a need to assess the **adaptiveness** of project management. For instance, are adaptations happening at the right time, what causes them, and so on.

Question 2: *Hans - you were involved in the process that led to the DAC evaluation criteria. Tell us about that experience and how these criteria came to be adapted so widely by the development community?*

Hans Lundgren: The DAC evaluation criteria have their origin in the DAC principles for evaluation which was one of the first tasks I was responsible for when assuming responsibility of the DAC Evaluation Network back in 1989. The criteria were then updated in 2002 with the Glossary of evaluation terms which was developed in collaboration with IEG. Both these processes involved extensive consultations and consensus-

building efforts, which were finally agreed to by all member countries and agencies. The criteria are part of a broader package of principles, guidance and standards developed by the DAC Evaluation Network. The criteria were conceived to help evaluation managers to reflect upon and structure the key question**s in an evaluation. I think one** reason behind their wide-spread use is that they are relatively easy to understand and to use when framing evaluation questions. Moreover, they relate to some key issues when assessing the success or failure of a programme.

Caroline: I agree with Hans that the criteria have been useful to shape overall questions about what we aim to assess. But, in practice I have seen too many evaluations that ask these questions without thinking. They use standardized – what made the program effective?, how efficient was the project?, etc. – without asking whether these questions are most important and useful. There are many other ways of asking questions that are more responsive to program managers, less jargonistic, and that will still lead to an assessment – or evaluative conclusion – of the relevance, effectiveness, efficiency, sustainability, and impact of programs that are evaluated.

Question 3: *Are all five criteria in the R/E/E/I/S framework still relevant? Is it time to review or replace all or some of them?*

Hans: Since your question asks if they are still relevant, I guess the criterion of relevance at least is still relevant! More seriously, I am personally open to look again at the criteria and see how they can be refreshed. But before throwing the adolescent out with the bathwater – the criteria have been in place for fifteen years now and not a baby anymore – we should reflect on what we can build on and the fact that since they have such a wide-spread use many consider them useful in practical work.

Caroline: True. It is not a matter of throwing the criteria out and starting all over. But, as evaluators we should take stock of how well they have worked and how they can be improved. I have made a number of suggestions in my recent blog series and we will take stock of all of the comments to think through the next steps.

Question 4: *Do you see some criteria as being more relevant for some types of programs/projects than others or are they applicable to most cases?*

Hans: The five criteria should not necessarily be used in all evaluations. The application of the criteria or any other criteria depends on the evaluation questions and the objectives of the evaluation. Furthermore, we have developed additional criteria in evaluating humanitarian aid and for peacebuilding activities in settings of conflict and fragility. I am in favour of a thoughtful application of these or other criteria not a mechanical application.

Question 5: *Revising the evaluation criteria is likely to be messy and difficult. Is it worth it? Can't we just work with what we have?*

Caroline: On the messiness of the process, Hans has a lot of experience in negotiating consensus among different parties. In addition to the challenges he points out, I would say that the tent has become bigger: there are more actors involved in development, which means there are more involved in evaluation. I would hope that a body like the OECD/DAC remains a standard setter and the legitimate convener of building consensus even with an enlarged group of players. But, in response to whether it is worth it? Yes, I do think so! The wide-spread use of the criteria demonstrates how important they – and the consensus around them – were. For evaluation – as a profession or practice – to adapt to modern times, it has to redefine itself periodically. Research into evaluation methods and their practical application are leading the way, but eventually we will have to update and redefine the norms.

Hans: It is true that developing and building consensus around internationally agreed norms and standards is a not a simple process, and I have spent years in my career on facilitating such consensus-building processes. It is not **only because** of the number of actors but because some countries and agencies may hold very firm positions. For instance, the DAC evaluation standards took three years to develop, test, revise and reach consensus on. An alternative to agreed, common approaches is of course that each agency and development bank develops their own criteria, norms and standards. However, this

would limit the possibilities of collaboration and reduce comparability.

Question 6: *One unintended consequence is that the criteria have potentially become somewhat of a straightjacket and lack the necessary flexibility. In other words, they foster a rigid structure* ***that produces*** *the same old reports spat out to the same old formula. Is this a fair criticism?*

Caroline: This critique is not new to me and often takes the shape of complaints about jargon that only evaluators can understand. I don't think this is a problem of the criteria as such, but has to do with their use, that is: the practice of evaluation. As I mentioned before, I have found evaluators – in many of the institutions that I have worked for – that have rigidly stuck to the criteria and were unable to use the criteria as the tool they were meant to be.

Hans: I am not sure which agency or development bank you have in mind when you say that they produce the same old reports spat out to the same old formulae. The application of the criteria has not blocked innovation as new methods and approaches have been developed during the last 15 years both for qualitative and quantitative evaluations. The criteria do not specify a specific method for evaluation but rather a way to help evaluators think about and structure the evaluation questions.

*Question 7: Are some criteria more important than others? Some have argued, for instance, that impact and sustainability matter more than efficiency, relevance, and effectivene*ss.

Hans: Which criteria are most important depends on the focus of the evaluation. There is obviously some interdependence between the criteria – if you get a number of positive effects it is also likely that your program was implemented effectively. One way of dealing with complexity and interdependence would be the merging of criteria which is mentioned in the blog series. At the same time, any changes need to be clear and practical in order to be applied.

Question 8: *In reviewing the criteria, how do we avoid the danger of being trapped into even more elaborate box-ticking approach to evaluation?*

Caroline: The problem that you raise is true, but not just for evaluation. I have seen this happen in many circumstances in the development field, and commented on the problem in evaluations I have written. I have not yet found the answer why this behavior occurs: is it the normal course of bureaucracies, or a natural response to ever more demanding agendas that ask too much for people to handle? At least initially, I do hope that we can keep the discussion of evaluation criteria and methods sufficiently "charged" to hold off on the more standardized responses or practices that you described as "box-ticking". In addition, my hope is that with an increasing number of evaluators who have dedicated their studies, research, and professional practice to evaluation, they will carry the banner that keeps renewing practices, including methods and criteria, to counter any risk of falling into stale routines.

Hans: As I am not in favor of a box ticking approach with the current set of criteria, I would not be in favor of a box ticking approach with a different set either.

Question 9: *There is a risk that incorporating the new criterion into evaluations will add complexity to what some already see as an already complex endeavor and entail a new learning curve. Is this where the development community should be spending its resources?*

Hans: In my view, to get wide spread use, any new criterion needs be to clear and not overly complex. I think there are other issues around "re-thinking evaluation" that the community needs to reflect on. An important issue is whether evaluation in its current form really provides policy makers with the evidence needed to make decisions on trade-offs between choices. Policy makers need to take decisions on alternative options, involving uncertainty and sometimes limited information. Perhaps evaluation work needs to become more exploratory in nature, rather than generating a historic record of accountability. Moreover, current evaluation and knowledge systems do not always function optimally and work remains to be done to improve use of evaluation findings and promote learning.

Caroline: Indeed, there are many things we need to work on, and that the criteria are only one of them. And while Hans is right that decision-makers

have to have evidence to weigh trade-offs between choices, this should not be limited to or even be primarily the responsibility of evaluation. In development banks, the appraisal of projects should include a comparison of the proposed solution with alternative options. Only: in practice that hardly ever happens. And, I do believe that an update to the evaluation criteria could incentivize the evaluation practice to address issues of importance to decision-makers. For instance, an evaluation that would evolve from an assessment of project relevance in its policy context to one that produces evidence whether the most impactful development challenge was addressed – as suggested in our blog series – would be a step towards answering questions in a more complex and uncertain world.

Question 10: *Complexity, agility, coherence, sustainability, and equity are examples of emerging areas in the area of evaluation. How are evaluators addressing these and other emerging issues?*

Hans: I think new approaches, new methods and new evaluation thinking are all to be welcomed. Evaluation research is leading the way and finding its way increasingly into practical evaluation work on such issues as complexity and equity for instance. But it would be good to see more experimentation and broader uptake of a variety methods. For instance, the use of big data in evaluation seems still to be in its infancy at least in development evaluation work. Further work on unintended effects would also seem to warrant more attention. Re-thinking evaluation however goes far beyond the discussion on criteria.

Caroline: Hans is right to say that rethinking evaluation goes beyond the criteria. As the past has shown: the criteria have incentivized a focus on certain aspects of development practice and can therefore be transformative if they are defined in line with current needs. That is not to replace the development, testing, and experimentation of new methods, but to stimulate and support these developments and keep with times.

Question 11: *Are we keeping up with trends outside the world of development evaluation? There is a vibrant and much larger universe of evaluation, beyond that of the development industry, that is continuously evolving and flourishing, and for which "rethink, reframe, revalue, relearn, retool and engage" is an embedded and ongoing process.*

Caroline: By all means: we are open to new ideas and improved practices. At IEG, we have hired a number of evaluation experts with the vision to upgrade our methodologies and evaluation practices. In addition, we are drawing on expertise and literature from any of the fields of evaluation to continuously grow.

Hans: I don't have the impression that the development evaluation field has gone stale and is inward looking. New articles in evaluation journals and books are being published constantly. And I am certainly in favor of promoting cross-fertilization from other areas.

Question 12: *Do the SDG's present an opportunity to reframe the evaluation dialogue and build the foundations for a more embracing, resilient, inclusive and sustainable world? What other drivers do you see as pushing the need to change?*

Hans: The Sustainable Development Goals as a vision for 2030 are certainly both an opportunity and a challenge. One lesson from the MDG era was that monitoring took the main role while evaluation was in the backseat. The implementation of the ambitious 17 goals, 169 targets, and the monitoring of 230 indicators certainly poses a number of challenges. From an evaluation perspective, I would like to see some more critical thinking: What is the theory of change? What about the assumptions in reaching the goals and targets? What steps need to be taken to enable evaluation to play a useful role in supporting implementation? A number of factors are driving change and disruption in our societies, including technology, violent extremism, competition between private firms and states - not only collaboration. Evaluators need to look outside the box.

Caroline: In addition, the SDGs include some targets on consumption patterns. If all countries aimed for consumption levels like those in OECD countries, the world overall would face considerable constraints and not achieve sustainability. Everyone needs to rethink consumption, including how we evaluate progress towards new consumption patterns. For instance,

the efficiency criterion asks whether project resources were used as efficiently as possible, but not whether the project (by design and in its final implemented state) contributes to wasteful consumption or sustainable consumption patterns. It is the most difficult part of the SDG agenda, is uncomfortable, and falls under no-ones mandate in particul**ar, which are ingredients for a** "forgotten" agenda that will be revived far too late, that is close to the 2030 target year.

Question 13: *Given the amount of interest that this topic has generated, how and where can stakeholders engage with you to build on the existing R/E/E/I/S framework going forward?*

Hans: The stakeholder group that I am most involved with is the DAC Evaluation Network which consists of some 40 evaluation departments from ministries, development agencies and banks. I believe there is an openness to discuss issues around "re-thinking evaluation". If a process of revisiting the criteria will be launched, it would be important to reach out widely to partners, civil society and evaluators in a consultative mode of engagement.

Caroline: Our first step will be to review the many comments and contributions we received on the blog series and then discuss with stakeholders, like Hans, whether and where to take this discussion. I agree with Hans that such a process would be open to wide-ranging consultation. ■

Biography

Caroline Heider is the Director General of the Independent Evaluation Group at the World Bank (IEG), a position she has held since 2011. Ms. Heider has dedicated the past 30 years of her career to evaluating the work of development and humanitarian organizations, transforming findings into lessons, and promoting innovative ways for institutions to apply the knowledge derived from evaluations toward accelerating development effectiveness. As a senior leader, Ms. Heider has a proven track record in leading change, strengthening institutions, and building evaluation capacity through testing and trying new methods to get to better evidence and greater insights. She has first-hand experience evaluating policies and programs in over 30 countries around the world.

Ms. Heider is a leading voice in the international evaluation community. She is a lifetime member of the International Development Evaluation Association (IDEAS) and a member of the American Evaluation Association. She chaired the Global Evaluation Advisory Committee of UN Women for the first years of its existence. In the past, she has been a member of the Australasian Evaluation Society and served a two-year term as vice-chair of the UN Evaluation Group.

Before IEG, Ms. Heider headed the Office of Evaluation at the World Food Program. She has also held leading positions in the evaluation offices of the Asian Development Bank and several UN agencies, including the International Fund for Agriculture Development, the UN Development Program, and UN Industrial Development Organization.

IEG
Independent Evaluation Group

WORLD BANK GROUP
World Bank • IFC • MIGA