Making M&E More 'Impact-oriented': Illustrations from the UN

Jos Vaessen, Oscar Garcia and Juha I. Uitto*

Abstract In international development, impact evaluation (IE) is becoming more and more an institutionalised practice. This article starts out by addressing the question of what institutionalisation of IE means and how it could work. Subsequently, the article explores common challenges in monitoring and evaluation functions in the UN system related to the supply of (and to a lesser extent demand for) evidence on impact. Rather than looking for solutions to these challenges in the practice of IE, the article explores the issue of how to improve non-IE monitoring and evaluation practices. On the basis of the identified challenges three categories of solutions are discussed: improving the quality of impact-related evidence at activity and project level, strengthening the causal logic underlying interventions, and strengthening the aggregation and synthesis of evidence. Finally, the article presents some illustrative examples of the latter two categories of solutions.

1 Introduction

In recent years, the widespread efforts to adopt results-based management[1] practices within the international development community have gone hand in hand with an increased interest in the effects of policy interventions. The widespread endorsement of internationally agreed upon development goals (e.g. the Millennium Development Goals, or MDGs) with corresponding indicators and targets at the impact level, increased pressures on public budgets allocated to development aid, new developments in technologies and systems of data collection, and repeated references to the paucity of available evidence on what works, are some of the key drivers behind this trend (e.g. CGD 2006; Jones *et al.* 2008). Correspondingly, over the last five years or so there has been a proliferation of impact evaluation (IE) exercises and the number of institutional actors involved in IE. In tandem with the increase in IE practices, debates on the methodology and practice of IE have flourished.

In the context of international development, governmental, non-governmental and inter-governmental organisations alike have taken steps to adopt and institutionalise IE practices in the context of their monitoring and evaluating (M&E) functions (see, for example, Jones *et al.* 2008; IEG 2009). This article starts out by addressing the question of what institutionalisation of IE means and how it could work. Subsequently, the article will explore common challenges in M&E functions in the UN system related to the supply of (and to a lesser extent demand for) evidence on impact. Rather than looking for solutions in the practice of IE, the article explores the issue of how to improve non-IE M&E practices. On the basis of the identified challenges three categories of solutions are discussed: improving the quality of impact-related evidence at activity and project level, strengthening the causal logic underlying interventions, strengthening the aggregation and synthesis of evidence. Finally, the article illustrates some of the solutions pertaining to these categories: using a theory-based review approach combined with a standardised rating system to develop insights about impact at portfolio level; improving the causal logic at higher levels of intervention using nested theories of change; and developing and using analytical tools to aggregate/synthesise patterns of impact at higher levels of intervention.

The article does not aim to provide a representative picture of the diversity in M&E functions and practices in the UN. Instead, it focuses on common challenges in the generation of evidence on outcomes and impacts, backed up

IDS Bulletin Volume 45 Number 6 November 2014
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

by examples from a UN perspective, and provides potential methodological solutions on how to address these challenges.

2 Pathways for institutionalisation of impact evaluation

Impact evaluations focus on the changes brought about by an intervention. While particular definitions of impact evaluation differ (for a discussion see White 2010; DFID 2012), there is a common understanding that impact evaluations focus on attribution,[2] i.e. the extent to which particular changes can be attributed to other interventions, taking into account other factors. In practice there are serious considerations of costs[3] and applicability of rigorous IE designs.[4] Consequently, coverage of portfolios of interventions at the level of a ministry, agency, country or region, will always inevitably be limited and biased towards particular types of interventions (Bamberger and White 2007). In order to properly address potential biases in coverage and other challenges in the application of impact evaluation, the latter should be embedded in a comprehensive M&E function. The institutionalisation of IE refers to a process which among other things includes a reflection about the when and how of IE within a particular organisational system. With respect to the latter, the Independent Evaluation Group of the World Bank has identified a number of characteristics of institutionalisation of IE:

- It is [a] country-led [process] and managed by a central government or a major sectoral agency;
- There is strong "buy-in" from key stakeholders;
- There are well-defined procedures and methodologies;
- IE is integrated into sectoral and national monitoring and evaluation (M&E) functions that generate much of the data used in the IE studies;
- IE is integrated into national budget formulation and development planning;
- There is a focus on evaluation capacity development. (IEG 2009: 14)

One can easily translate the abovementioned principles to a UN organisational context: stakeholder buy-in, well-defined procedures and methodologies for IE, IE integrated in and supported by institutional M&E functions, IE connected to decision-making processes, and attention to capacity development. Institutionalisation is important for several reasons (for a discussion see Gaarder and Briceño 2010). When embedded in an organisational M&E function, the planning of IE exercises is likely to improve (taking into account considerations of coverage, priorities and value for money), the interaction with institutional stakeholders is likely to be more efficient and there is a higher potential for use of the findings (e.g. in terms of institutional learning and/or informing strategic decisions). Moreover, the potential linkages between IE and the design and planning of interventions as well as links with other M&E practices may lead to better data as a basis for IE exercises. These arguments are equally valid whether IE exercises are carried out by programme units or by independent evaluation offices. In both cases, IE is enabled by the availability of reliable data and monitoring systems, which in turn emphasises the need to design programmes with this in mind.

In practice, one can discern several possible pathways for institutionalising IE. The Independent Evaluation Group of the World Bank (IEG 2009) describes three different pathways: *ad hoc* impact evaluation (e.g. Colombia), sector-initiated series of impact evaluations (e.g. Mexico) and whole-of-government-initiated strategy for IE (Chile). The first pathway emphasises the exemplary effect of conducting a number of high-profile impact evaluations. The principle is the following. As a result of one or two successes, i.e. impact evaluations influencing the policy and/or public debate, stakeholder demand for IE increases. Consequently, this gradually leads to a formalisation of IE in terms of a more systematic planning of IE, the development of norms and standards, and a formalisation of processes of communication and use of impact evaluations. The second pathway emphasises the role of a series of impact evaluations in a high-profile area of work (e.g. social protection and conditional cash transfers) supported by donors and institutional champions in the country, resulting in products that have policy value and consequently provide the basis for further formalisation and scaling-up of IE practices. Finally, the third pathway concerns the systematic planning of IE as part of a broader M&E function linked to national planning and budgeting processes. As a result of several

iterations and continuous institutional learning on the planning, execution and use of impact evaluations the system can be further improved.

Both IEG (2009) and Gaarder and Briceño (2010) emphasise the importance of at least two key factors for successful institutionalisation: the role of champions and the linkages between IE and existing M&E functions. A very particular manifestation of the institutionalisation of IE is the example of the USA where in some policy areas rigorous IE is a mandatory prerequisite for informing the continuation or increase in public funding for programmes. Epstein and Klerman (2012) convincingly argue that mandatory IE is not always cost-effective as the absence of effects can also be the result of design flaws or implementation failures. In order to detect these one does not need costly and rigorous IE but more conventional evaluation techniques such as evaluability assessments or process evaluations. This strengthens the argument that IE should be part of a broader M&E strategy.

There are other arguments that emphasise the need for linking IE to existing M&E functions. For example, rigorous IE requires adequate data; without *ex ante* data available, IE may be less credible and more expensive. Second, many types of interventions are not amenable to statistical counterfactual analysis or other types of quantitative techniques that are used to address the attribution challenge. Increasingly, qualitative methods are available for addressing attribution and other methodological challenges in IE (see DFID 2012 for an extensive debate). However, rigorous impact evaluations (including a broader range of IE approaches than those based on (quasi-) experimental designs) are costly and there are always challenges of applicability of particular methods to certain interventions given the questions of interest to stakeholders[5] (Bamberger and White 2007). Consequently, there is a need for coherent planning and use of different M&E tools in order to be able to address biases in coverage[6] and develop a reliable perspective on impact at the level of portfolios of interventions.

Given the importance of linking IE to existing M&E functions one could raise a rather fundamental question. If the objective is to increase the quality of the evidence base on how and the extent to which an organisation is making a difference in the world, what would be the best strategy? Options include promoting the practice of IE (i.e. see the discussion above on the three pathways) and/or strengthening institutional M&E functions towards becoming more 'impact-oriented'. We purposely refrain from providing a clear definition of what the latter term exactly means. Ideally it would encompass changes in the decision-making environment (the institutional context of IE), the data availability and methodological improvements in non-IE M&E practices. In a way, the strengthening of existing M&E functions towards becoming more 'impact-oriented' can be conceived of as an additional pathway to institutionalisation of IE. Through changes in existing (non-IE) M&E practices one can, for example, improve the conditions (e.g. through the identification of knowledge gaps, the articulation of causal linkages, the availability of data) for conducting relatively low-cost and high value for money IE (Kusek and Rist 2004). In the remainder of this article we will make the case that this fourth pathway actually presents a more 'cost-effective' way forward in terms of providing evidence on impact (across portfolios of interventions) rather than just promoting the (costly) practice of IE. In practice, both non-IE practices and IE should be promoted, yet the emphasis should be first and foremost on the former.

3 Challenges for strengthening impact evaluation practices in the UN

Since the second half of the previous decade, there has been a notable increase in impact evaluations in development, driven by on the one hand the demand for evidence on impact by donors and on the other by the realisation that the evidence base on what works and what does not in international development has been weak (CGD 2006). New institutional players such as the Bill and Melinda Gates Foundation, the International Initiative for Impact Evaluation and the Poverty Action Lab as well as established institutions such as the World Bank and DFID, to name a few, have been at the forefront of promoting the practice of impact evaluation in international development. Through these institutions and other initiatives such as the Network of Networks on Impact Evaluation, the demand for and supply of IE evidence has also increased within the UN.

There are no recent comprehensive records on the prevalence of impact evaluation in the UN system. In 2009, in a survey among member organisations of UNEG, nine (out of 28 members who responded)[7] reported the practice of IE in their organisational systems (UNEG 2013). Some of the frontrunners in the application of IE are the International Fund for Agricultural Development (IFAD), World Food Programme (WFP), United Nations International Children's Fund (UNICEF), the Global Environment Facility (GEF) and a few others. Even in some of the smaller organisations nowadays one can find IE practices. In the latter case, they are usually one-shot exercises supported or requested by particular donors. Most IE exercises within the UN are not relying on quantitative counterfactual approaches.[8] Many are based on non-experimental quantitative approaches, theory of change approaches, and other qualitative methods. Combinations of qualitative methods or qualitative and quantitative methods are quite common.

Since 2009, there certainly has been an increase in IE practices across the UN, but at the same time it has become clear that IE is not a prevalent practice in most organisations within the UN system. Overall it can be concluded that most UN agencies have limited demand as well as budgets for IE (UNEG 2013). This has implications for the institutionalisation of IE. While in some UN organisations this process may be strengthened through a (gradual) introduction of IE exercises (e.g. WFP), at the same time many UN organisations are confronted with resource and other constraints (see discussion further on in this section) which inhibit the widespread and cost-effective use of IE. Consequently, it is worthwhile to explore the potential of the fourth pathway (identified above), strengthening the 'impact-orientedness' of M&E functions, as a basis for strengthening the evidence base on impact.

On the basis of combined experiences of the authors as well as recent documentation, the following key challenges for IE in the UN system can be identified:

- There are a number of constraints on the demand side for IE evidence which among other things imply that financial resources for IE are scarce.[9]
- The use of evaluations in multilateral environments (as in other public sector environments) and correspondingly the demand for evaluative evidence is highly diverse and it is quite challenging to capture the different institutional processes that affect use and demand. A recent literature has pointed out several challenges in the context of use and demand of evaluative evidence in international development (Pritchett 2001; Bamberger 2010; McNulty 2012).
- There are substantial differences between organisations within the UN system and across the board one cannot yet speak of a culture of learning from evidence on impact. The institutional arrangements for evaluation vary considerably, although there appears to be a general trend towards institutionalising evaluation as an independent function often with a primary purpose towards accountability.
- IE[10] can most easily be applied to clearly delineated interventions affecting clearly delineated target groups. While there is a lot of work in the UN that fits these criteria, for example work that directly affects communities and individuals (e.g. humanitarian work, vaccination programmes, improving curricula in primary schools), a lot of the work conducted by the UN is at (inter)national institutional levels (e.g. policy advisory work, global knowledge products, normative work, etc.). It is much more difficult to rigorously establish the impact of the latter types of interventions (White and Phillips 2012) than in the case of the former.
- Many interventions (co-)implemented by the UN (independent of the level and nature of intervention) are complicated: they are multi-actor, multi-stranded (multiple intervention components) and multi-level (e.g. international, national, local). While sometimes these interventions can be deconstructed into evaluable components, sometimes they cannot. In the case of the former there is always the danger of bias (see previous bullet point).[11]
- Decision-makers in the UN system, for example at the level of the governing bodies, are not necessarily interested in the impact of a particular intervention. Evaluating the impact of one intervention can be methodologically quite challenging. Yet, the generation of evidence on impact at higher

levels adds two additional challenges: the challenges of aggregation and alignment.[12] While there is an extensive literature on addressing challenges in attribution and there are signs that this knowledge is increasingly internalised by evaluators and programme staff across the UN, expertise on how to address the methodological challenges of credibly aggregating and synthesising evaluative evidence is much less widespread.

4 Examples of potential solutions: towards more 'impact-oriented' M&E functions

In this section, we briefly illustrate some of the potential solutions to addressing the challenges discussed above, drawing from actual cases.

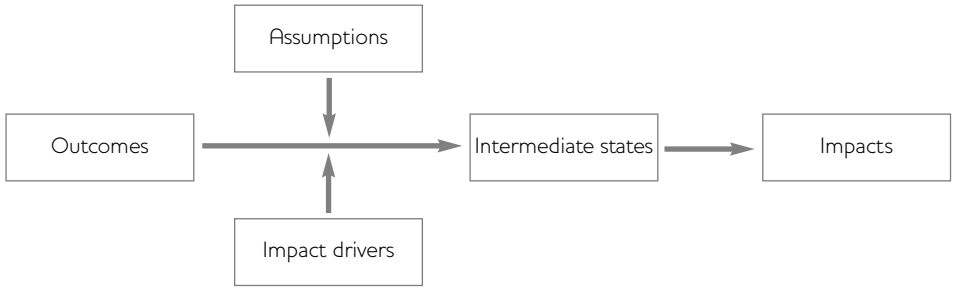
In general, it is useful to distinguish between three categories of solutions: improving the quality of impact-related evidence at activity and project level, strengthening the causal logic underlying interventions, and strengthening the aggregation and synthesis of evidence. The first category refers to the sources of data and the quality of data collection and analysis at activity and project level. There is much to be said about this broad topic and we will not discuss it in detail here due to reasons of scope and space. One key element that falls in this category and which is quite pertinent in the context of the UN (but again also for other public sector organisations) concerns the respective roles of self-assessment (and reporting) and (external) evaluation. In principle, self-assessment is quite adequate for activity and output delivery analysis, as these aspects can be relatively easily observed or captured and are in the control sphere of the intervention. By contrast, outcomes and impact are more difficult to assess due to challenges in attribution. It is here that external evaluations have a comparative advantage. However, despite the fact that programme staff who undertake the self-assessments often do not have the time, resources, data or incentives to provide reliable and unbiased assessments of outcomes (and impact), self-assessment continues to play an important role in outcome and impact analysis and reporting in the UN. External evaluations which rely on dedicated resources, expertise and an (often) independent analysis of (progress towards) outcomes and impact[13] may feed into results-reporting but this is not necessarily always the case and even if it is the case it may not happen systematically. Consequently, improving the evidence base on outcomes and impact in a UN context requires taking a closer look at the comparative advantages and respective roles of self-assessment and external evaluation and, consequently, changing the system in favour of the latter.

The second category concerns the use of theories of change in the planning, monitoring and evaluation of interventions. There is significant scope for a more widespread use of theories of change in the planning and M&E of interventions within the UN and beyond. By clearly articulating the causal linkages between intervention activities, outputs, outcomes and impact, including assumptions about the influence of external factors, one creates a better basis for credible and comprehensive data collection and analysis on processes of change towards impact. This principle lies at the heart of the rationale for theory-based evaluation (Chen 1990; Rogers *et al.* 2000) and it is useful to distinguish between two levels of analysis. First, the use of theory-based evaluation at the level of a 'simple' intervention (e.g. an activity or project). As the purpose and application of theory-based evaluation at the level of a 'simple' intervention have been widely discussed and are relatively well known it will not be further discussed here. Instead, we will illustrate further on in this section an example of the second level of analysis. Multi-level interventions (programmes, portfolios, strategies) encompass multiple intervention activities at different levels, all of which can be deconstructed into specific theories of change. These different theories can subsequently be 'nested' inside an overall theory of change of the multi-level intervention.

Finally, there is a category of solutions that relates to the aggregation and synthesis of evaluative evidence. While there are specific evaluation approaches such as multi-site evaluations and (systematic) reviews that would feature in this category (see Vaessen and Van den Berg 2009), all of the examples discussed next involve a specific approach to aggregation.

Our aim is not to present a comprehensive and exhaustive set of solutions (which in fact would be impossible to achieve). Instead, through some examples we will elucidate the second and third

Figure 1 **Generic model of a theory of change used in the Review of Outcome to Impact**



Source GEF (2009).

category of solutions (the use of nested theories of change and approaches to strengthen the aggregation and synthesis of evidence). As stated above we will not further discuss the category of solutions relating to improving the quality of impact-related evidence at activity or project level.

We focus on three examples which will illustrate the following solutions:[14]

- Strengthening the aggregation and synthesis of evidence: the use of a review method and standardised rating system to assess the quality of causal analysis at project level, assess the extent to which a project is (likely) to contribute to (expected) processes of change and report on impact at higher levels of analysis (Global Environment Facility, GEF);
- Strengthening the causal logic of interventions and strengthening the aggregation and synthesis of evidence: the use of nested theories of change to make sense of the effects of a complicated multi-level, multi-stakeholder policy intervention (United Nations Educational, Scientific and Cultural Organization, UNESCO);
- Strengthening the aggregation and synthesis of evidence: the use of a custom-made review method to identify patterns of causes and effects occurring at country level (United Nations Development Programme, UNDP).

Example 1: the Review of Outcome to Impact model (ROtI) in the GEF

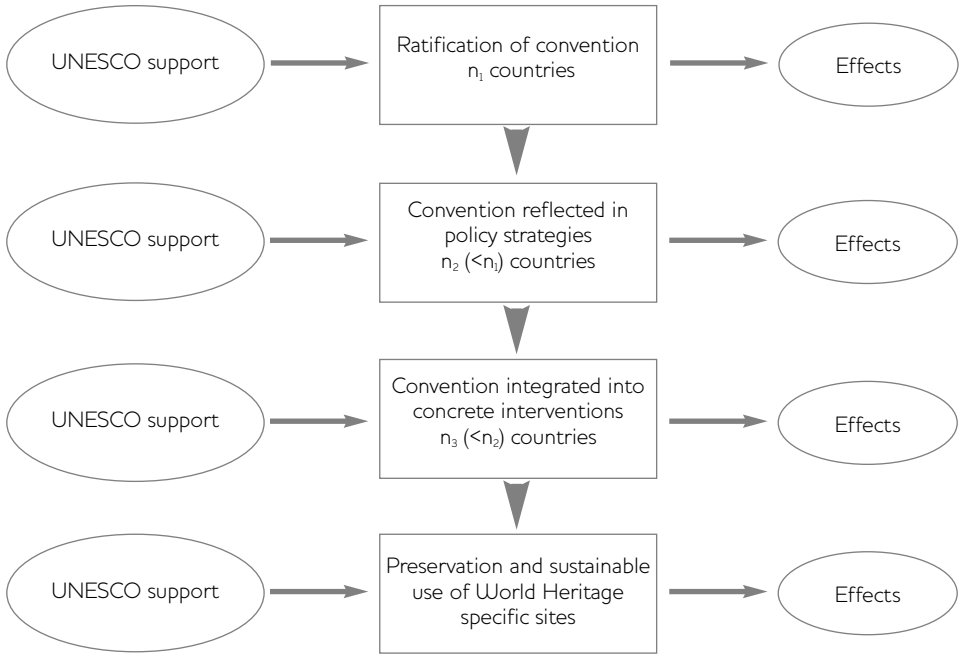
As part of the fourth overall performance study, the GEF Evaluation Office developed an approach to assess the overall performance and progress towards impact of GEF projects. The Review of Outcome to Impact (ROtI) approach has been developed as an alternative to time-consuming and expensive full impact evaluations. Two variants of the ROtI have been developed, one based on desk research only and another one with limited additional primary data collection.

Broadly, the approach follows three stages (see GEF 2009). Stage one develops the theory of change model (ToC) of the project. First, the intended outcomes and impact of the project are identified. Second, the project's logical framework is reviewed to understand how the project's outputs are expected to contribute to outcomes (behavioural changes) and impact (societal and environmental change resulting from behavioural changes). Third, the outcome to impact pathways are elaborated, thereby distinguishing between intermediate states, assumptions and external drivers (see Figure 1; see also GEF 2009). In case of the desk ROtI, the main sources of information are the initial project document and monitoring reports.

In stage two, evidence on the different components (see Figure 1) of the ToC is identified. In case of the desk ROtI evidence is mainly found in the project's terminal evaluation report. On the basis of the search and identification of evidence and fitting it into the ToC model, some level of assurance on the likelihood of achievement of outcomes and progress towards impact can be obtained. In the last final stage, ratings are attributed to the project, which are directly linked to the extent of evidence available on the different components of the project's ToC.

Advantages of this approach are that one can aggregate ratings across projects and areas of work, providing some level of assurance on progress towards impact at portfolio level. Some

Figure 2 Simplified structure of the theory of change of the 1972 convention showing the different levels of intervention



Source UNESCO (2013).

of the drawbacks are: gaps in evidence on outcomes and impact; limited assurance on attribution; and, limited use for learning purposes due to the aggregation of data that do not reveal the underlying very heterogeneous contexts and interventions.

Example 2: Articulating the nested theories of change of the UNESCO Convention concerning the protection of the world's cultural and natural heritage

In the preparatory phase of an evaluation of UNESCO's standard-setting work on culture, the Evaluation Section of UNESCO's Internal Oversight Service in consultation with programme staff reconstructed a series of theories of change relating to specific culture conventions. One of these conventions was the 1972 Convention on the Protection of Cultural and Natural Heritage.

A convention is an example of a multi-level intervention. UNESCO and implementing partners intervene at different levels (international, national, programme) to promote and implement the principles of the convention with different stakeholders and target groups across the world. Figure 2 shows the structure of the overall theory of change of the convention[15] with the different levels of intervention. At each level of intervention different intervention activities are implemented with different constellations of stakeholders. To illustrate:

- Among other things, at the international level UNESCO[16] promotes the ratification of the convention and facilitates the dialogue between member states which ratified the convention. The latter among other things contributes to shared priorities and increased collaboration among member states.
- Among other things, at the national level UNESCO helps raise the awareness of decision-makers (e.g. parliamentarians, senior civil servants) on the importance of the convention and how it relates to national policy initiatives. As a result, principles of the convention are included in national strategies and plans.
- Among other things, at the policy and programme implementation level UNESCO assists member states in designing interventions in alignment with achieving the objectives of the convention. As a result, among other things, institutional capacities in line ministries and government agencies are built on how to support the preservation and sustainable use of natural and cultural heritage.

- Among other things, at the level of individual heritage sites, UNESCO assists member states in the assessment of threats to preservation. This and other types of activities directly contribute to the protection of world heritage.[17]

The complexity of evaluating the convention becomes clear when one takes into account that at each level of intervention there are different intervention activities with different stakeholder configurations in specific contexts. Moreover, the scale and number of activities can vary widely. A simple example is the reduced number of countries where UNESCO is active in a particular period in time[18] when one moves down from the international level to the site-specific level. A first step in the evaluation process is to develop the overall framework of expected causal linkages connecting activities and outputs to expected processes of change. Subsequently, for each of the levels, as depicted in Figure 2, more specific nested theories of change can be developed as guidance for causal analysis.

Consequently, in order to assess the overall relevance, effectiveness (and even impact) of the work of UNESCO carried within the framework of this convention, one can synthesise information according to the logic of different theories of change at different levels in different contexts, including the data on activities, outputs (and outcome indicators) in the theory (some of the quantitative data can be simply aggregated, qualitative data can be synthesised). Subsequently, by 'nesting' all of the theories and corresponding empirical data back into the perspective of the overall theory of change one can develop the overall perspective.

The advantage of this approach clearly lies in its potential to make sense of a complex multi-site, multi-level intervention. Drawbacks are the time and resources that need to be devoted to developing and agreeing upon the different theories of change and using it as a basis for rigorous data collection. Ideally, a multi-site (impact) evaluation should be conducted to ensure homogeneity and rigour in data collection and analysis.

Example 3: UNDP meta-review of country programme evaluations

As part of the independent evaluation of the UNDP Strategic Plan 2008–2013 (SP) the Evaluation Office of UNDP commissioned a meta-review of evaluations to assess UNDP's development contribution against stated goals in the areas of (1) poverty reduction and MDG achievement; (2) democratic governance; (3) crisis prevention and recovery; and (4) environment and sustainable development.

The evaluation conducted an overall analysis of UNDP effectiveness, efficiency, sustainability and cross-cutting performance against explicit sub-criteria. Evidence was available from three types of evaluation: the 'Assessment of Development Results' (ADRs) reports (the country level evaluations of UNDP performance), thematic evaluations at the global level and project evaluations. Thirty-one ADRs and nine thematic evaluations that were completed between 2010 and 2012, and which were managed and quality assured by the Evaluation Office (EO), were selected for inclusion in the analysis. Project evaluations that were commissioned by country offices were not included in the analysis due to limited time available.

The first part of the analysis consisted of a systematic assessment and rating approach to assess whether UNDP performance was highly unsatisfactory, unsatisfactory, satisfactory or highly satisfactory on the different criteria, using clear guidance on the attribution of ratings.

The second step, which was more innovative, looked at the factors explaining these outcomes. Many evaluations do not explicitly identify factors which contributed to (or impeded) the achievement of development outcomes and in which contexts. The evaluation used a *realist logic* approach to identifying the major factors explaining observed variations in performance, and combined it, to the extent feasible, with Qualitative Comparative Analysis (QCA).

In practice this approach meant adopting a specific way of extracting information from the ADRs and thematic evaluations. All these evaluations, to varying degrees of explicitness, developed theories on what were the major contextual factors and programme characteristics mainly responsible for the specific outcomes found. A realistic evaluation approach (see Pawson and Tilley 1997) was used to model these theories into CMO (context-mechanism-outcome) configurations that were systematically analysed using QCA. Converting the extracted

statements to a format that QCA can handle, entailed conceptualising contexts and mechanisms into mathematical sets. Each case needed to be scored on context and mechanism conditions, using different scales that were dichotomous (yes/no, presence/absence) or fuzzy (membership scores, e.g. 0, 0.1, 0.2, 0.25, 0.33, 0.6, 0.75, 1). A membership score of 1 indicated full presence/membership while a membership score of 0 indicated complete absence (fully out).

ADRs were a particularly reliable source for comparative data because they were similarly structured and contained roughly the same kind and amount of information. The statements made in the ADRs needed to be analysed in terms of whether they resembled CMO statements. A database of such statements, as extracted from each ADR, was constructed, and the statements systematically compared. At the end of this process a number of CMO patterns were identified on combinations of conditions associated with similar SP outcomes. Similarity of outcomes was considered first along specific SP outcomes – that is by comparing similar scores within the same specific outcome. Later, these were merged if CMO patterns seemed to occur across different specific SP outcomes.

The study aimed at limited or 'middle range' generalisation, rather than universal generalisation. The objective was to discover a limited number of different paths to success (or failure) that are at work in different contexts. An important challenge in carrying out a meta-synthesis review of evaluations of this type is maintaining consistency among evaluation findings under each sub-criterion for each evaluation. This turned out to be equally challenging for the ADRs and the thematic evaluations. To manage the risk of deviations in information extraction, protocols were developed and an external review of all completed assessments was undertaken and any anomalies in classification were discussed with the individual reviewer. Finally, making configurations comparable was critical for ensuring the quality of synthesis. Once the level of comparability was satisfactory, synthesis was carried out in compliance with recognised QCA practices (see, for example, Befani, Ledermann and Sager 2007).

A major advantage of the applied approach was the potential to generate generalisable patterns of regularity across studies. Some of the remaining challenges concerned the heterogeneity and depth of information available from different studies.

5 Conclusion

This article discussed the thesis that in order to strengthen the evidence base on impact, it is not sufficient, even not cost-effective, to 'merely' promote the practice of impact evaluation. Instead, the marginal utility of improving existing (non-IE) M&E tools towards becoming more 'impact-oriented' is higher. Not only would this lead to better answers on the extent to which interventions are making a difference, it would also ameliorate the basis for identifying strategically important gaps in the evidence base and opportunities for cost-effective IE exercises.

Impact evaluations are situated at the level of specific interventions. While there are useful techniques available such as (systematic) review with qualitative and/or quantitative synthesis to aggregate and synthesise evidence from single impact evaluations, building up a broad base of impact evaluations across an organisation's portfolio is costly. By using tools such as nested intervention theories, as illustrated in the example, and better targeting data collection and analysis towards the causal impact pathways at different levels of intervention, one can strengthen the evidence base and at the same time get a better picture of where in-depth IE may add the highest value for money.

A related argument for focusing on (non-IE) M&E tools concerns the need for evidence at aggregate level. In the UN system, the demand for evaluative evidence mostly related to programmatic results at a higher aggregate level beyond individual projects. This need for knowledge is equally important for accountability of the results achieved as well as learning for improved programming. Two examples in this article illustrate how such evidence on impact can be generated at the aggregate level.

In sum, strong impact-oriented M&E functions both provide evidence on the performance of the organisation (and progress towards impact) and allow for evaluation units to design and conduct impact evaluations based on solid data and on those parts of the portfolio where they are most needed and useful.

Notes

* The ideas and opinions expressed in this article are those of the authors and do not necessarily represent the views of UNESCO, UNEP or UNDP and do not commit these organisations.

1 'Results' is a generic term and in the UN has been defined as referring to changes at output, outcome and impact levels (UNDG 2011).

2 Some prefer the term 'contribution' to emphasise a reality of a confluence of factors influencing change. In essence there is no tension between the concepts. Attribution analysis always involves a reflection (and appropriate methodological responses) about other factors that influence a particular change.

3 The authors' experiences with impact evaluations conducted in the context of multilateral and bilateral cooperation suggest that independent of methodological design, impact evaluations are costly exercises. OED (2005) estimated the cost of impact evaluations as between US$200,000 and US$1,000,000. More recent estimates also point at amounts (significantly) above US$100,000 with substantial variation, depending on the nature and scope of the evaluand among other things.

4 The applicability of (quasi-)experimental designs is more limited than other designs such as theory-based approaches (see Bamberger and White 2007). Independent of design, the concept of rigour is associated with data (analysis) requirements, which in turn have implications for applicability.

5 In principle, the impact-related questions of interest should guide the choice of methods and not the other way around.

6 This refers to the idea that some types of interventions are more likely to be subject to IE than others, which can lead to biased coverage of a portfolio of interventions by IE. For example, a common critique against the promotion of randomised controlled trials has been that it generates a bias in coverage as only particular interventions are amenable to this type of IE design.

7 UNEG has 43 institutional members.

8 A substantial number of scholars and practitioners equate these approaches with 'rigorous'. While this is a long discussion (see DFID 2012), one could argue that these approaches are in principle strong on the attribution issue (internal validity). External validity, however, is harder to judge (Bickman and Reich 2009). Bamberger (2010) reckons that, although there are no hard statistics available, it is quite likely that rigorous IE designs (in the sense of quantitative counterfactual analysis) are only used in perhaps 10 per cent of ODA impact evaluations.

9 Yet, due to pressures on financial resources and pressures from member states, the demand for and use of evaluative evidence on impact is increasing.

10 Especially quantitative counterfactual analysis.

11 'A reality that often has to be faced in IE is that there is a trade-off between the scope of a programme and strength of causal inference. It is easier to make strong causal claims for narrowly defined interventions and more difficult to do so for broadly defined programmes. The temptation to break programmes down into sub-parts is therefore strong; however, this risks failing to evaluate synergies between programme parts and basing claims of success or failure on incomplete analysis' (DFID 2012: ii).

12 See White (2003) for a discussion. In the context of IE, aggregation refers to the issue of how information at lower levels of intervention can be meaningfully aggregated. Alignment concerns the relevance of information available at lower levels of interventions in terms of supporting claims on the achievement of higher-level objectives.

13 Keeping in mind that most evaluations are not impact evaluations.

14 Examples of other methodological innovations include the opportunities for using 'Big Data' in analysing development processes and effects or the use of (systematic) review exercises to distil generalisable lessons on what works and why under what circumstances across different intervention contexts. More generally, and this goes for project-level and (to some extent) higher level interventions, in cases where IE is inappropriate, other evaluation methods such as real-time, action-research oriented and formative evaluations can be used to fill gaps in evidence on processes of change and impact (DFID 2012).

15 The overall theory of change of the 1972 convention that was developed in collaboration with programme staff is fairly complex and elaborate and not included here.

16 And implementing partners.

17 The very act of including a site in the world heritage list can generate significant effects such as increases in tourist income and direct and indirect local employment effects.

18 In principle, all countries (even those that have not ratified the convention, at the national level could be covered by UNESCO assistance if circumstances would demand it, but for a given period of time UNESCO's activities are mostly concentrated on subsets of countries (e.g. where new proposals for sites are being prepared, where sites are under threat, etc.).

References

Bamberger, M. (2010) 'Institutionalizing Impact Evaluation. A Key Element in Strengthening Country-Led Monitoring and Evaluation Systems', in M. Segone (ed.), *From Policies to Results: Developing Capacities for Country Monitoring and Evaluation Systems*, New York NY: UNICEF

Bamberger, M. and White, M. (2007) 'Using Strong Evaluation Designs in Developing Countries: Experience and Challenges', *Journal of Multidisciplinary Evaluation* 4.8: 58–73

Befani, B.; Ledermann, S. and Sager, F. (2007) 'Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application', *Evaluation* 13.2: 171–92

Bickman, L. and Reich, S.M. (2009) 'Randomised Controlled Trials: A Gold Standard with Feet of Clay?', in S.I. Donaldson, C.A. Christie and M.M. Mark (eds), *What Counts as Credible Evidence in Applied Research and Evaluation Practice?*, Thousand Oaks CA: Sage Publications

CGD (2006) *When Will We Ever Learn? Improving Lives Through Impact Evaluation*, Report of the Evaluation Gap Working Group, Washington DC: Center for Global Development

Chen, H.T. (1990) *Theory-Driven Evaluation*, Beverly Hills CA: Sage Publications

DFID (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, Working Paper 38, London: Department for International Development

Epstein, D. and Klerman, J.A. (2012) 'When is a Program Ready for Rigorous Impact Evaluation? The Role of a Falsifiable Logic Model', *Evaluation Review* 36.5: 375–401

Gaarder, M. and Briceño, B. (2010) 'Institutionalisation of Government Evaluation: Balancing Trade-offs', *Journal of Development Effectiveness* 2.3: 289–309

GEF (2009) *The ROtI Handbook: Towards Enhancing the Impacts of Environmental Projects*, Washington DC: Global Environment Facility

IEG (2009) *Institutionalizing Impact Evaluation Within the Framework of a Monitoring and Evaluation System*, Independent Evaluation Group, Washington DC: World Bank

Jones, N.; Walsh, C.; Jones, H. and Tincati, C. (2008) *Improving Impact Evaluation Coordination and Uptake*, scoping study commissioned by the DFID Evaluation Department on behalf of NONIE, London: Overseas Development Institute

Kusek, J. and Rist, R.C. (2004) *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*, Washington DC: World Bank

McNulty, J. (2012) 'Symbolic Uses of Evaluation in the International Aid Sector: Arguments for Critical Reflection', *Evidence and Policy* 8.4: 495–509

OED (2005) *OED and Impact Evaluation: A Discussion Note*, Operations Evaluation Department, Washington DC: World Bank

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, Thousand Oaks CA: Sage Publications

Pritchett, L. (2001) *It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation*, Working Paper, Cambridge MA: Kennedy School of Government

Rogers P.J.; Hacsi, T.A.; Petrosino, A. and Huebner, T.A. (eds) (2000) *Program Theory in Evaluation: Challenges and Opportunities*, New Directions for Evaluation 87, San Francisco CA: Jossey-Bass

UNDG (2011) *Results-Based Management Handbook*, New York NY: United Nations Development Group

UNEG (2013) *Impact Evaluation in UN Agency Evaluation Systems – Guidance on Selecting, Planning and Managing Impact Evaluations*, UNEG Impact Evaluation Task Force, New York NY: United Nations Evaluation Group

UNESCO (2013) 'Draft Structure of the Theory of Change of the World Heritage Convention', unpublished working document, Internal Oversight Service, Paris: UNESCO

Vaessen, J. and Van den Berg, R. (2009) 'Evaluative Evidence in Multi-Level Interventions: The Case of the Global Environment Facility', in O. Rieper, F.L. Leeuw and T. Ling (eds), *The Evidence*

Book: Concepts, Generation and Use of Evidence, New Brunswick NJ: Transaction Publishers

White, H. (2010) 'A Contribution to Current Debates in Impact Evaluation', *Evaluation* 16.2: 153–64

White, H. (2003) 'Using the MDGs to Measuring Donor Agency Performance', in R. Black and H. White (eds), *Targeting Development: Critical Perspectives on the Millennium Development Goals*, London: Routledge

White, H. and Phillips, D. (2012) *Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework*, Working Paper 15, New Delhi: International Initiative for Impact Evaluation