THE EES NEWSLETTER

EVALUATION CONNECTIONS

Robert Picciotto Editor. **Ole Winckler Andersen** Associate Editor. **Tereza Zvolská** Editorial Assistant.

EES EUROPEAN EVALUATION SOCIETY

DECEMBER 2015

ARTICLES – NEWS – EVENTS

CONTENT

A MESSAGE FROM THE VICE PRESIDENT

Riitta Oksanen

Dear EES members:
The International Year of Evaluation 2015 has been unique in many ways. Rather than a centrally planned process with coordinated implementation it has celebrated evaluation through close to a hundred individual events around the world. The organisers of each of these events selected themes relevant to their constituencies and responsive to the distinctive demands of their operating context. This bottom-up approach shaped a versatile agenda of evaluation issues that reflects the highly diverse priorities of active (and often enthusiastic) evaluation practitioners and advocates. The lighting and passing of the evaluation torch from event to event has symbolized the shared goals connecting all events.

The same spirit of crowd-sourcing of energy and ideas characterized the approach of the EvalPartners movement from its inception. EvalPartners, the initiator of the International Year of Evaluation 2015, defines itself as an innovative partnership to enhance the capacities of Civil Society Organizations (CSO) to influence policy makers, public opinion and other key stakeholders so that public policies are based on evidence, and incorporate considerations of equity and effectiveness.

The overarching objective of the initiative is to enhance the capacities of CSOs to engage in a strategic and meaningful manner in national evaluation processes, contributing to improved country-led evaluation systems and to policies that are equity-focused and gender equality responsive. While the focus of EvalPartners has been on CSOs in general and the voluntary organisations of professional evaluation (VOPEs) in particular, the network of supporting institutions has brought together several UN agencies, multilateral financing institutions and national governments.

The International Organisation for Cooperation in Evaluation (IOCE) provides a home base for the initiative. The initiative got kick started in the first EvalPartners Global Forum held in Chiang Mai (Thailand), on December 2–6, 2012. The culmination of the International Year of Evaluation will take place during the second Global Forum to be held in Kathmandu on November 23–25, 2015. The forum will be hosted by the Parliament of Nepal. This is another sign of the mobilising power of the movement: parliamentarians committed to use evaluation have started organising themselves in networks under the EvalPartners umbrella.

The Global Evaluation Forum in Nepal will take stock; summarise the results of on-going EvalPartners' consultations and synthesize the conclusions of all International Year of Evaluation events. It is also expected to adopt

a Global Evaluation Agenda (2016–2020). The EvalPartners consultations have been conducted within an evaluation capacity development framework adopted in Chiang Mai in 2012 that reflects the complex and multiple perspectives of EvalPartners' stakeholders worldwide.

The framework includes three levels:

(i) a *national enabling environment* that provides a cultural context for institutional and individual initiatives buttressed by political will and evaluation policies.

(ii) an *institutional context* comprised of independent evaluation units, adequate budgets, institutional policies and ethical guidelines – key pillars of evaluation capacity.

(iii) a capabilities framework at the *individual evaluator* level that outlines the knowledge, skills and dispositions needed to deliver high quality, impartial evaluation services – the lynchpin of evaluation excellence.

The framework also identifies the importance of strong supply of and demand for evaluation. "Supply" refers to the capability of the professional evaluation community to provide sound and trustworthy evaluative evidence. "Demand" refers to the capacity and willingness of policy makers and senior managers to request sound and trustworthy evaluative evidence with the aim of using it in strategic decision-making processes in the public interest.

EES has been an active participant in the Global Evaluation Agenda consultation process. It has hosted a Vice Presidential blog on the EES web site. It has co-sponsored a seminar organised by the Finnish Evaluation Society (FES) in Helsinki about *The Future of Evaluation: a Global Perspective through a Nordic Lens"* (September 17th, 2015). Its EES Emerging Evaluators thematic working group organised a virtual conference on *"The Future of Evaluation… for the Future"* on September 19th 2015. It has partnered with UNESCO, OECD and France's evaluation society (SFE) to set up a Conference about *Making Effective Use of Evaluations in an Increasingly Complex World* held on September 30th 2015 in Paris. Finally under the impetus of the EES Sustainability Thematic Working Group a symposium about *"Evaluating the Sustainable Development Goals – New Challenges for Research, Policy and Business"* was held on October 28th 2015 in Vienna, in partnership with the Institute for Managing Sustainability (Vienna University of Economics and Business) and in collaboration with the Austrian Development Agency (ADA).

The conclusions of all EES sponsored events and consultations have illuminated many of the issues already evoked by the first drafts of the Global Agenda. The critical role of a favourable enabling environment and of parliaments was reaffirmed. Special emphasis was put on the fundamental importance of *values* in evaluation as an integral element of transparency in governance and an amplification of citizens' voices in the democratic process.

At the institutional level maintaining independence while avoiding isolation continues to challenge the evaluation community and capacity development is widely perceived as a priority in pursuit of evaluation excellence. In this context EES participants have emphasised the value of working together to gain benefits from international networking and have stressed VOPEs' key role in knowledge sharing and professional development with special attention to the needs and aspirations of emerging evaluators.

With respect to individual evaluators capacities EES has embraced its responsibilities in the professionalization of evaluation. The definition of fixed competency thresholds tested at various levels of mastery is bound to be controversial. On the other hand, the EES capabilities framework focused on self examination helps provide precious insights and it offers a foundation for pertinent feedback. Once facilitated by voluntary peer review it would enhance professional development and expand evaluators' abilities to function and produce valuable work in diverse contexts. Inclusiveness of all evaluators and new ideas is another basic principle that is guiding implementation of the Voluntary Evaluator Peer Review pilot.

Finally and most importantly the EES consultation process has highlighted three gaps in the current global evaluation agenda:

- The increasingly rapid pace of decision making in an increasingly volatile operating environment and the challenges that result for evaluation methods and processes
- The expanding role of the private sector in society and the implications for evaluation governance, models and approaches
- The on-going information technology revolution and its likely impact on data generation, collection and analysis in evaluation.

These are among the issues that will to be explored at the 12th Biennial Conference in Maastricht.

Riitta Oksanen
EES Vice President

AN EDITORIAL: DOES PROCESS MATTER TO EVALUATION QUALITY AND USE?

Ole Winckler Andersen

Methodological and institutional issues have been at the center of recent evaluation debates. Extensive discussions have also taken place regarding evaluation professionalization and of the need for capacity building and training.

By contrast there has been little attention to evaluation processes. Limited information exists about how evaluation processes influence evaluation products and outcomes. While the need for continuous interaction between evaluators and commissioners throughout the evaluation process is widely acknowledged the motivations, incentives and strategies of the various actors are poorly understood.

Conversations with evaluators and commissioners indicate that evaluation processes may play a more significant role than usually envisaged, but while many recognize the important role of evaluation processes, they may not agree on the character of this role. This leads to a number of questions related both to how various factors influence the evaluation process, and how the role of the evaluation processes should be assessed. It is also relevant to ask a basic normative question: are current evaluation processes 'fit for purpose' in diverse and evolving governance contexts? Securing answers to such questions call for more analyses of how evaluation processes are conducted and what impact they have on decision making and the public interest.

The papers in this issue of Evaluation Connections show that different perspectives and analytical approaches can be applied when trying to assess the role of evaluation processes. The articles that follow also indicate the complex character of evaluation processes. Several of them suggest that more research is needed, e.g. through process tracing techniques.

The first paper by Robert Picciotto introduces a number of alternative evaluation governance models and discusses based on three criteria (moral hazard; transaction costs/ information asymmetries; and responsiveness to the public interest) the strengths and weaknesses of each model. The paper thus provides a useful analytical framework for assessment of evaluation processes within alternative governance systems.

The following two papers by Anna Paterson and Chris Barnett highlight the important role played by evaluation processes from two different perspectives. Paterson's paper discusses the implications of the size of the market for evaluation and of the type of actors active in this market. Barnett's paper points at a number of ways an evaluation process can be distorted and recommends improvements in commissioning and management of evaluations. Both papers call for further research in evaluation processes

Agathe Devaux-Spatarakis's paper is based on a study of a number of RCTs conducted in France 2006–13. The paper provides an account of how different interests have influenced the implementation of the RCTs, and the stark conclusion of the paper is that the different interests and expectations of the actors have had significant implications for the final evaluation products and their use.

The paper by Per Øyvind Bastøe and Øyvind Eggen discusses the feasibility of using evaluations to guide budget allocations. The two authors identify several challenges and conclude in particular that beyond the current focus on Terms of Reference and choice of evaluators timing is important for potential use of evaluation findings. Equally more emphasis should be placed on what to evaluate.

In his paper Sebastian Lemire calls for a stronger role for consequential validity in evaluation. Consequential validity can serve as a valuable guide to evaluation process design by bridging approaches to evaluation based on methodological rigor and use. According to the paper this would also strengthen the awareness of how specific evaluation designs and processes may lead to adverse consequences.

Finally Mathew Hall's paper presents three evaluation 'logics' – scientific, bureaucratic and learning – and argues that the 'learning logic' is best suited for evaluations of third sector organizations. Here again it emerges that evaluation approaches and processes should be specifically designed to meet the distinctive needs of users in diverse contexts.

EVALUATION IN ORGANIZATIONS: THE TRICKY RECTANGLE

Robert Picciotto

All evaluation governance models require effective quality assurance arrangements including a strict focus on ethical standards. Poor quality evaluations are dangerous. Unethical evaluations do harm. Misguided recommendations can be disruptive and counterproductive. Naive participatory methods can lead to evaluation capture. The wrong indicators can distort incentives.

This helps explain why the evaluation literature has focused on the relationship between the evaluator and the evaluand and why it has neglected the role that evaluation plays in organizations. This observation led Bastiaan de Laat to conceive of a *tricky triangle* model connecting the evaluator, the commissioner and the evaluand (2014). But his trilogy excluded the ultimate beneficiary – the citizen. Bringing the citizen out of the cold implies an alternative model of evaluation governance: *the tricky rectangle* (Figure 1).

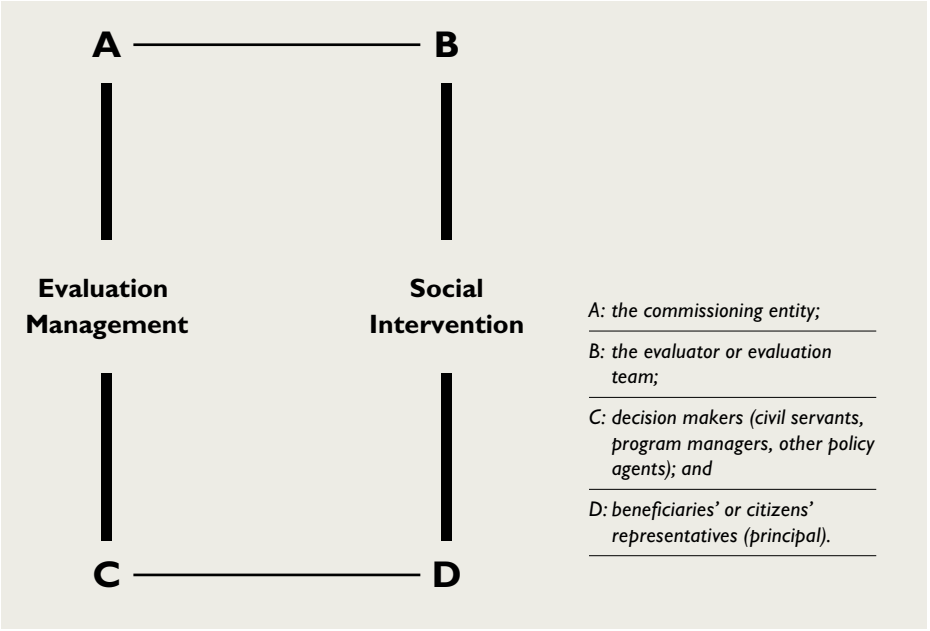

Figure 1: The "tricky rectangle" of evaluation governance.

The entities located at the four corners of the rectangle interact in diverse ways. No single configuration model holds sway over all others. Tradeoffs must be struck in order to tap the diverse benefits that evaluation can confer on society. In a nutshell, judicious linkages between the four corners of the triangle have the potential to improve incentives (less free riding), reduce transaction costs and align principal and agent goals thus minimizing conflict of interest and moral hazard.

The most frequent evaluation governance model combines A and C, i.e. decision makers commission the evaluation and the evaluator (B) is a hired gun. This *market based* configuration reflects the growing role that vested interests play in the evaluation world. It turns evaluation into a private good. Under this option evaluators operate as management consultants. The configuration is characterized by high transaction costs (contracting, oversight) and it severely limits the extent to which the evaluator has sufficient autonomy to protect the interests of the citizen (D).

Under this model evaluations are frequently commissioned to improve the image of the organization; to help one internal faction prevail over another; to rebut public criticism; to delay needed change; or simply to comply with a statute rather than to engage in the evaluative process in a principled way.

If A, B and C are one and the same – i.e. if the commissioner, the evaluator and the decision maker are combined (*self evaluation*) – the same democratic deficit as in the previous option prevails but information asymmetries and transaction cost are lower and evaluation processes are timelier and more efficient. In fact high quality self-evaluation has many advantages: it improves policy design and monitoring and its self generated findings are more likely to be owned by decision makers. However self evaluation lacks objectivity.

The *enhanced self evaluation* model conflates B and C – the commissioner (A) contracts the program manager (C) to act as evaluator (B). This variation enhances the independence of the self evaluation configuration and improves its responsiveness to the public interest to the extent that A reflects the interests of D, the citizen.

If the commissioner (A) is the same as the evaluator (B) but distinct from the decision maker (C) independence is substantial but legitimacy hinges on the extent to which A is aligned with D or if to the contrary whether it represents an interest group. This is an *advocacy evaluation model.*

Finally the *independent evaluation* configuration keeps A, B and C at arm's length. In this configuration A is mandated to represent the principal (D: the citizen) and commissions B to evaluate the work of the agent (C: the program manager or policy maker). This is how

parliamentary commissions and the General Accountability Office report operate.

Similarly at the World Bank, an Independent Evaluation Group reports to executive directors who represent governments and their citizens. This configuration protects evaluation independence without incurring isolation. The model minimizes information asymmetry and relies on the checks and balances that distribute authority between board and management so as to protect evaluation independence.

External evaluation cannot be equated with evaluation independence. The judgment of external evaluators may be impaired or threatened if their services are retained by the managers in charge of the activities being evaluated: fee dependence is a major threat to the integrity of the evaluation process. By contrast, internal evaluation units funded and controlled by a supreme oversight authority are protected from management interference through the checks and balances associated with democratic governance arrangements.

This said, independence on its own does not guarantee evaluation quality: relevant skills, sound methods, adequate resources and transparency are also required. Independent but incompetent evaluations can be misleading, disruptive and costly. On the other hand, evaluation quality without independence lacks credibility. This is why persons and entities that have reason to fear the outcome of an evaluation will frequently throw doubt on its independence.

The independent evaluation model can be combined with the enhanced self evaluation model thus tapping synergies. Just as accounting benefits from auditing self- evaluation benefits from independent evaluation. Program managers and policy makers often lack evaluation skills. They may have different interests and concerns than program beneficiaries. They may succumb to leaps of faith that lead to faulty or excessively risky decisions. They may be tempted to select evidence that supports their pre-conceptions.

Professional oversight of self evaluation by independent evaluation (facilitated by the enhanced self evaluation feature) improves the effectiveness of self- evaluation in ways similar to those that make internal auditing useful in the judicious implementation of administrative policies and procedures.

	Moral hazard	Transaction Costs/ Information Asymmetries	Responsiveness to the public interest
Market oriented	High	Medium	Low
Independent	Low	Medium (for internal units) High (for external units)	High
Self Evaluation	High	Low	Low
Enhanced Self Evaluation	Medium	Medium	Medium
Advocacy Evaluation	High	Medium	Depends on nature of the advocacy group (faction vs. civil society)

Table 1: Quality of alternative evaluation governance models.

The relative advantages and disadvantages of the five configurations are displayed in the table above (Table 1).

No single configuration is ideal. A tailor made approach to the design of evaluation functions in organizations is desirable. The objective should be to minimize transaction costs, minimize moral hazard and ensure that the public interest is not sacrificed at the altar of evaluation utilization. From a democratic evaluation perspective the most appropriate governance configuration combines independent and self evaluation.

Reference

de Laat, B. (2014). The Tricky Triangle, *Connections,* March.

WHO DELIVERS EVALUATION CONTRACTS AND HOW MUCH DOES IT MATTER?

Anna Paterson

In order to consider the effects of actors' incentives and constraints within a political system, the actors need to be identified. I began researching this article by trying to map the actors that are contracted to deliver UK government-commissioned development evaluations. I was struck by two things during these first steps that ended up consuming the entire article. First, in line with the established view of the development evaluation market, the range of actors involved in the delivery of evaluation was not large and there appeared to be a dominance of certain types of contractors. Secondly, when it came to searching for literature on the implications of the size of the market and of the types of actors within it, this appeared still to be under-researched.

The article was written in the context of a new approach in the UK Department for International Development (DFID) since 2011 whereby the evaluation function was embedded in spending units, aimed at increasing the quantity, quality and coverage of evaluations, including rigorous Impact Evaluations.[1] It followed there would be an increase in evaluation contracts to be delivered by external agents. DFID procures many large evaluations through the Global Evaluation Framework Agreement (GEFA), which allows access to a panel of pre-qualified suppliers. Meanwhile, in other UK government departments there are a variety of approaches to commissioning and managing external evaluation and many departments also conduct at least some of their evaluations in-house. (NAO, 2015)

There are different types of potential evaluation providers in international development, including Northern and Southern universities and for-profit and non-profit actors. Many of the actors that deliver evaluation contracts also deliver implementation contracts, but evaluation Terms of Reference do demand mechanisms to declare, prevent and manage direct conflicts of interest. Where before the 1980s, universities and non-profit organisations dominated development contracts, since the 1980s, for-profit firms have played a much bigger role. (Dickinson, 2005; Huysentruyt, 2011) There has been something of a blurring of the lines between universities, NGOs and for-profit contractors, since academics will often be part of teams set up by for profit firms or NGOs, many universities have set up profit generating consultancy arms, and some non-profit development think-tanks are quite academic in their work. I found little published work investigating differences and similarities in the range of contractors for international development evaluation compared to contractors commissioned to evaluate policies and interventions for other government departments, or comparing types of contractors used by different development agencies.

Since the new approach to evaluation, DFID has published two Annual Evaluation Reports in 2012–13 and 2013–14 which list, and link to, evaluations published in those years. My brief analysis suggested that of these 52 evaluations completed and published[2] from 2012–14 (25 in 2012–13 and 27 in 2013–14), 31 (59.6 %) were led by consultancy firms, 12 (23.0 %) were conducted by individual consultants or teams of consultants with no clear affiliation to a larger firm,[3] 4 (7.7 %) by universities, and four by other types of agent.[4] DFID's procurement processes and Evaluation Strategy are committed to widening the market of evaluation suppliers by 'procuring evaluations from a range of providers located in the global North and South'. (DFID, 2014a) However, DFID's own analysis has shown that there has been limited competition for the majority of bids in the GEFA and a dominance of Norther actors. (DFID, 2014b) In spite of DFID's and other development partners' attempts to build the development evaluation market the supply of service providers has remained thin both in the UK and worldwide (DFID, 2014c)

What do the thinness of the international development evaluation market and the types of suppliers within it mean for evaluation processes or outcomes? The answers to these questions appear unclear based on current research, data and analysis. DFID has certainly considered the implications of over-reliance on Northern evaluators and is committed to building the capacity of Southern evaluators. However, there are other possible implications that deserve attention.

A number of studies have identified relationships between the different incentives of different types of contractors and performance in government contracts. One study that modelled these relationships for implementation contracts in international development found that organisational identity of contractors had important effects on bidding processes, on the flexibility to deliver a specified design versus adherence to the contractor's own prior preferred course of action, and on dimensions of contract performance and cost. (Huysentruyt, 2011) There is room for further investigation of the role of contractor type in evaluation contracts specifically.

The dimensions of interest in such an investigation might include, for example, the highly valued, but complex and often poorly defined, aim of independence. (Mayne, 2012) There have been studies that look at the

1. These new evaluations would complement and provide data for the Independent Commission on Aid.
2. One further evaluation was completed but not published for reasons of sensitivity.
3. Although one of these was produced by an individual consultant using material produced by a larger consultancy firm.
4. Three by other development partners, and one apparently by a DFID call-down resource centre.

trade-offs between managerial independence and operational use and usefulness of evaluation units in development agencies. (Foresti, 2007) It is likely that similar trade-offs apply between types of evaluation contractors. Indeed, studies and analyses of evaluations in other UK government departments have suggested that academic evaluators may be seen as more independent but more aloof and less able to provide timely findings to maximise utilization, whereas consultant evaluators might be more delivery focussed and closer to practitioners but may also be less critical. (Salisbury et al., 2011; Dixon, 2015) This is echoed in some development literature suggesting that the effects of factors such as the need to win future work and the repeated interaction between, and mutual dependence of, consultants and commissioning units should at least be explicitly considered and unpacked. (Copestake and Williams, 2014)

Another area of potential consequence involves the possible relationship between the types of evaluation contractors in the market and the ability to deliver different evaluation methods. Studies of evaluations in other departments have suggested relationships between contractor type and method, with management consultants being more associated with shorter 'formative evaluations' and university led evaluations with more robust multi-centre trials over several years. (Salisbury et al., 2011) Many of the skills required to deliver robust impact evaluations are more likely to be found amongst specialised academics and indeed DFID supports Impact Evaluations through a number of dedicated initiatives such as the International Initiative for Impact Evaluation (3ie) and the Abdul Lateef Jamil Poverty Action Lab (JPAL). But as noted above, in international development there has been a blurring of the lines dividing academics from other types of evaluators and this may have implications for the way academically rigorous evaluation methods are delivered.

This short reflection has yielded more questions than conclusions, but it does suggest that the characteristics of those who deliver evaluations may be important in ways that we have not fully researched or understood. Further research and analysis in this area would help to build upon a small but growing body of work that considers the micro politics of relations between donors and other stakeholders. (Copestake and Williams, 2014; Andersen, 2015)

References

Andersen, O. W. (2014). Some thoughts on Development Evaluation Processes, of Befani, B., Barnett, C. and Stern, E. (eds.) *Rethinking Impact Evaluation for Development*, IDS Bulletin, Volume 45, Number 6, UK: Wiley Blackwell, pp. 77–84.

Copestake, J. and Williams, R. (2014). *'Political Economy Analysis, Aid Effectiveness and the Art of Development Management'* Development Policy Review 32 (1).

DFID Evaluation Strategy 2014–19 (June 2014a), p. 10.

DFID Evaluation Strategy 2014–19 (June 2014c) p. 10.

DFID (February 2014b) *Rapid Review of Embedding Evaluation in UK Department for International Development*, p. 53.

Dickinson, L. A. (2005). *'Government for Hire: Privatising Foreign Affairs and the Problem of Accountability under International Law'* (pp. 135–235). William and Mary Law Review Vol. 47.

Dixon, A. (2015). *'We need Critical Friends and Robust Challenge, Not Aloofness and Separation'* (p. 221). Blog-post on the Policy Innovation Research Unit blog, London School of Hygiene and Tropical Medicine http://blogs.lshtm.ac.uk/piru/2015/05/12/we-need-critical-friends-and-robust-challenge-not-aloofness-and-separation/.

Foresti, M. (2007). *A Comparative Study of Evaluation Policies and Practices in Development Agencies*, Overseas Development Institute (ODI) and Agence Francaise de Development (AFD).

Huysentruyt, M. (2011). *Development Aid by Contract: Outsourcing and Contractor Identity*, London School of Economics & Stockholm School of Economics.

Mayne, J. (2012). *'Independence in Evaluation: The Role of Culture'* in Barbier, Jean-Claude & Hawkins, Penny (eds.) (2012). Evaluation Cultures: Sense-making in Complex Times, Comparative Policy Evaluation, Vol 19.

National Audit Office (2013). *Evaluation in Government*, Cross-government Report, NAO, UK.

Salisbury, C. et al. (2011). *'Making the Most of Evaluation: A Mixed Methods Study in the English NHS.'* Journal of Health Services Research & Policy Vol 16 No. 4, p. 221.

AN EVALUATOR'S PERSPECTIVE: IMBALANCES OF POWER IN EVALUATION DESIGN AND IMPLEMENTATION

Chris Barnett

The importance of the relationship between 'commissioner' and 'evaluator' is often underplayed within the evaluation process – and yet these very dynamics can fundamentally affect the rigour and use of the evaluation findings. Poor evaluation design is frequently attributed to technical inadequacies and capacity constraints; evaluators not choosing the right methods, or not having the capacity to apply them rigorously. Indeed much of the debate over the past decade has focused on methods and ways to improve rigour (e.g. Savedoff et al., 2006; White, 2009), and with far less attention on the politics of the evaluation process. In this brief article, I share some observations on why we need to think more carefully about the different interests and asymmetries between commissioner and evaluator.

Apart from the well-documented drive towards experimental and quasi-experimental methods, even work that has argued for a more *appropriate* use of methods (e.g. Stern et al., 2012) has tended to focus more on the conceptual and methodological challenges – and remains largely silent on the trade-offs that take place during the evaluation process itself. While others have taken up this challenge in the past – such as the link between methods, resource and time constraints (most notably Bamberger et al., 2006) – there is still a lack of emphasis on how the real life process of commissioning, plus the different stages of an evaluation, are subject to all kinds of distortions. This is less from a methodological, managerial or practical point of view, but rather one where distortions are driven by political interests, institutional incentives and the workings of a dysfunctional evaluation system (Michaelowa and Borrmann, 2006; Andersen and Broegaard, 2012; Andersen, 2014).

These are not minor considerations. Indeed, the disconnection between 'client' and 'supplier' should not be underplayed as this can significantly affect evaluation findings. For example, a meta-analysis of CGIAR rates-of-return studies (Walker et al., 2008) notes that *external* consultants tended to be overly optimistic (or *insiders* were overly pessimistic) about the impact of technological change (Alston et al., 2000). Clearly, it was more than just methodology that led to this systematic bias. In another example, Walker et al (2008) cites how donors tend to be willing to fund impact assessments on 'hot topics' and that this can result in under-resourcing assessments in more conventional areas. Such biases arise not from individual choices that occur in isolation, but through the institutional processes, interests and incentives embedded within them.

More often than not, the objectives and interests of the different actors in the evaluation process are not fully aligned. This is often exacerbated by traditional procurement and managerial practices. For example, many procurement models used for commissioning evaluations cause fundamental information asymmetries at each stage of the design process. Whereas the literature and evaluation guidelines assume that evaluation design occurs as a one-off 'event' wholly under the control of the evaluator, this is untrue. The reality is that most (especially donor-led) evaluations are procured through competitive tendering processes where design occurs at a number of stages with different actors; each with different interests, and each artificially separated from the dialogue needed to resolve these differences.

Typically, there is a pre-tender stage which involves developing the specification in the form of a Terms of Reference. But here, the commissioner *works in isolation* from the evaluator let alone the evaluand. Then, there is a proposal stage, where the evaluator *works in isolation* from the commissioner. It is often not until a third (and subsequent stages, if there is extensive negotiations), that that the commissioner and the evaluator (plus other stakeholders) finally come together in a process where the commissioner holds all the cards. By then, it is usually too late to alter many of the fundamental parameters and choices made around evaluation questions, design choices, and resources – even though this is the first time that the different interests of the commissioner, implementer and evaluator are brought together and surfaced. Throughout the process there is a perverse incentive for the evaluator not to walk away or raise fundamental issues.

Current commissioning processes lead to wild mismatches of client-supplier expectations. In one example I know, the evaluator's estimates of the resources necessary to implement the evaluation were some fifteen times the actual resources available! Clearly the commissioner's expectations in this particular case were out-of-kilter with the realities of the assignment. More often than not however, differences of expectation and interest are concealed – with design considerations overlooked in order to reach timely agreement about price or scope leading to inevitable frustrations and misunderstandings about delivery and quality later on in the evaluation process.

Over and above the problems associated with mismatched expectations, lack of transparency can lead to systematic bias. Where the commissioner is in charge of the programme being evaluated, the imperative

1. This seems counterintuitive, as one might have expected insiders to overestimate rates of return. The suggestion in Walker et al. (2008) is that in this particular case, outsiders were less familiar with the interventions and thus overoptimistic in their assessments, perhaps in part due to incentives to extend their employment. In other circumstances, it may be that outsiders are more pessimistic, but this example only serves to further reiterate the need for more understanding of the systematic bias that results from different commissioner-evaluator arrangements.

of robust evaluative evidence may conflict with incentives for continued programme support and self-legitimisation. This can lead to overoptimistic 'success claims', the selective use of evidence and a reluctance to offer recommendations to cancel programmes or change them.

For these reasons, there is a need to focus on better transparency and alignment of the differing interests of commissioner and evaluator, including dealing with information asymmetries at different stages of the evaluation process. The "blueprint" mentality whereby evaluations can be fully specified upfront, procured and then simply implemented is not conducive to evaluation excellence. This is all the more so given the growing complexity of interventions, the prevalence of partnerships, and the dynamic and fragile contexts in which they operate. Evaluations are rarely uniform or standardised and can't be procured in the same way as one might do for office chairs, or a financial audit. There are too many unknowns, too much complexity, and too many trade-offs that need to be decided during the course of the evaluation – but more than that, evaluations are political, with different stakeholders having vested interests in the eventual findings.

It is therefore time for a fundamental rethink. Firstly, around the way evaluation is commissioned and procured so as to deal with the information asymmetries between commissioner and evaluator. Indeed, we are starting to see some progress with new ways of operating, such as early market engagements, an increased use of evaluability assessments, plus two-stage evaluations (where the first inception stage helps draft or revise the terms of reference for the second stage). These are all attempts to reduce the antagonistic relations between commissioners and evaluators, and resolve market dysfunctions where commissioner and supplier's expectations are mismatched. Secondly, we need some new thinking on more agile modes of management that could be applied to the evaluation process – i.e., providing structures and processes whereby the commissioner and evaluator can make adjustments as expectations and interests become better aligned over the course of an evaluation while respecting the independence of the evaluator.

And lastly, and perhaps where we have seen least progress so far, is the need for renewed thinking about the governance arrangements for evaluations: How do we best mediate different interests, and in such a way that it protects and enhances the credibility, independence and the usefulness of findings? And, how can this consider stakeholder interests beyond the commissioner-evaluator relationship?

References

Alston, J. M., Chan-Kang, C., Marra, M. C., Pardey, P.G. and Wyatt, T.J. (2000). *A meta-analysis of rates of return to agricultural R and D ex pede herculem?* Research Report 113, IFPRI, Washington DC, USA.

Andersen, O. W. and Broegaard, E. (2012). The political economy of joint-donor evaluation, of *Evaluation* 18(1), Sage Publications, pp. 47–49.

Andersen, O. W. (2014). Some thoughts on Development Evaluation Processes, of Befani, B., Barnett, C. and Stern, E. (eds.) *Rethinking Impact Evaluation for Development*, IDS Bulletin, Volume 45, Number 6, UK: Wiley Blackwell, pp. 77–84.

Bamberger, M., Rugh, J., and Mabry, L. (2006). *Real World Evaluation: Working under Budget, Time, Data, and Political Constraints*, Sage Publications.

Michaelowa, K. and Borrmann, A. (2006). Evaluation Bias and Incentive Structures in Bi- and Multilateral Aid Agencies. *Review of Development Economics*, 10(2) pp. 313–329.

Savedoff, W. D., Levine, R. and Birdsall, N. (2006). *When will we ever learn? Improving lives through impact*, Evaluation Report of the Evaluation Gap Working Group, Center for Global Development, Washington, USA.

Stern, E., Mayne, J., Befani, B. Stame, N., Forss, K. and Davies, R. (2012). *Developing a broader range of rigorous designs and methods for impact evaluations*, Final report, DFID, UK.

Walker, T., Maredia, M., Kelley, T., La Rovere, R., Templeton, D., Thiele, G. and Douthwaite, B. (2008). *Strategic Guidance for Ex Post Impact Assessment of Agricultural Research*, Prepared for the Standing Panel on Impact Assessment CGIAR Science Council, June 2008.

THE DIFFICULT ENCOUNTER BETWEEN EXPERIMENTAL EVALUATION PROCESSES AND STAKEHOLDERS' INTERESTS

Agathe Devaux-Spatarakis

The growing use of Randomized Controlled Trials (RCTs) has elicited an abundant literature regarding their operational relevance, their ability to address causation and their compliance with ethical standards. (Donaldson and Christie, 2005) This brief article raises another concern: the effect of RCT processes on the relationship between evaluators and stakeholders, e.g. project managers and policy actors. What happens when RCT protocols are deployed in the field?

This article addresses this issue. It is grounded on empirical data from an unpublished PhD research on the use of RCTs in France between 2006 and 2013. It demonstrates that RCTs can be studied as a social institution. From a constructivist perspective RCTs are a set of rules, resources and roles that are shaped in turn by the reactions of those who are subjected to RCT practices rules. The research summarized here included fifteen cases studies of RCT interactions with different stakeholders and the impact on evaluation results. Three distinct phases structured the analysis: inception, implementation and use.

The first focus of our inquiry was the *inception* phase of RCTs. In France, the majority of evaluators practicing RCTs are academics. They are eager to adopt a method which only started to be used in France in 2006. It promises unambiguous tests of micro-economic theories following scientific protocols and enables evaluation practitioners to publish in highly ranked scientific journals. Hence, evaluators are mostly pursuing a scientific interest in the conduct of RCTs which enjoy the additional advantage of benefiting from generous public funding.

Most of the experiments reviewed for this article were funded by the French government through bidding processes open to joint applications by project managers and evaluators. Project managers had an incentive to become involved since the RCT dimension helped to trigger funding for their projects and enhance the likelihood of future funding. Equally evaluators adopted the RCT approach since it increased their chances of being selected and many of their academic colleagues perceived it to be a gold standard.

Although some project managers had prior associations with experimental evaluators and shaped their project jointly with them from the start, many others had to shop around in order to find evaluators qualified to carry out RCTs and convince them to team up. In many cases the managers and their evaluator partners were unfamiliar with the detailed technical requirements of RCTs so that the project designs failed to include enough beneficiaries to achieve statistical validity and/or were structured in ways that failed to strictly control the implementation of the project.

The *implementation* phase brought forth conflicts related to different expectations about the conduct of experiments. These were revealed as two prerequisites of RCT protocols were implemented: (i) random allocation of the treatment and (ii) control of the treatment itself. First and foremost programme managers, being used to tailor the social intervention to fit the distinct needs of intended beneficiaries, viewed compulsory random allocation of a standard package of services as a denial of their expertise and a hindrance to the quality of service delivery.

Therefore, rather than complying with RCT protocols they often chose to provide access to the treatment to all eligible members of the public and exercised flexibility in treatment of individual cases. This approach, while ethical, undermined the scientific validity of the experiment.

Another prerequisite of statistical validity for RCTs hinges on identifying a sufficiently large number of potential beneficiaries and mobilizing them to observe the RCT protocols. Fulfilling this condition proved hard to do since not all individuals allocated to the "test" group chose to accept the treatment or comply with its protocols. Thus in many experiments, even if the original design included a sufficiently large cohort, a substantial part of the public turned down the opportunity to participate in the experiment. This was not a surprise to project managers, as disadvantaged citizens are often difficult to reach and new interventions often need adjustments and time to induce their participation.

Yet another dimension of conflict between RCT principles and the reality of field practice had to do with control of treatments. Although most interventions were properly framed at the outset many either evolved as the experiment took place or were changed to meet the distinctive needs of the agents who delivered the intervention. In a nutshell, for a wide variety of reasons, none of experiments examined by this research followed the scientific protocols planned by the academic evaluation team. Of course the discrepancy between experimental protocols and the ways experiments actually unfold in the field has been underlined in the literature even for the most celebrated examples in the use of RCTs for social research. (Faulkner, 2014; Rodrik, 2008)

Inevitably the interpretation of results from flawed experiments proved problematic. Rigorous scientific standards were hard to meet due to small cohorts and contamination between the two groups. In the cases under study, results were only produced after a long process of analysis and sophisticated statistical treatment. More often than not they were considered as partially valid, and in need of confirmation through further experiments. Thus the interventions failed to generate valid and useful policy recommendations.

Project managers did not benefit from the approach since the experiments did not allow them to make use of results in real time so as to adapt their practices in light of experience.

They also realized that they had unrealistic learning expectations from RCTs as they were expecting to get a comprehensive understanding of the programme in context, whereas RCT results were only focused on one aspect of the intervention that could be accurately assessed by this method. (Rodrik, 2008)

Positive results led to generalization only when the intervention did not disturb current implementation practices. Negative results led to the end of the programme when politicians lost interest in the intervention due to new political priorities, or because in the first place, the aim of the experiment was to demonstrate the inefficiency of the intervention and win an ideological debate.

In conclusion our findings suggest that the RCT 'scientific' method can only be strictly applied where policy actors have a deep understanding of RCT requirements. Furthermore learning expectations should be made coherent and interventions should be shaped accordingly. Unless these preconditions are met major questions as to the capacity of RCTs to accurately account for the impact of public action in the real world and to feed policy learning will linger.

References

Deaton, A. (2010). Instruments, Randomization, and Learning about development, *Journal of Economic Literature*, 48, pp. 424–455.

Donaldson, S., Christie, C. (2005). The 2004 Claremont Debate: Lipsey versus Scriven. Determining causality in program evaluation and applied research. Should experimental evidence be the gold standard? *Journal of Multidisciplinary Evaluation*, 3, pp. 60–77.

Faulkner, W. (2014). A critical analysis of a randomized controlled trial evaluation in Mexico: Norm, mistake or exemplar?, *Evaluation*, 20(2), pp. 230–243.

Rodrik D. (2008). The new development economics: We shall experiment, but how shall we learn, *Brookings Development Conference*.

EVALUATIONS AS BASIS FOR BUDGET DECISIONS

Per Øyvind Bastøe and Øyvind Eggen

When the current Norwegian government took office in 2013, an expressed intention was to '*carry out systematic evaluations that will directly affect budget decisions, and facilitate full transparency regarding the scope, implementation and impacts of Norwegian development policy*'[1].

In this article we will – from our position in the Evaluation Department – explore if this is a realistic expectation, with a particular emphasis on evaluations that directly affect budget decisions. As we will describe below, past experiences are not very encouraging when it comes to deliver on this intention. However, some lessons may be drawn regarding what it will take to meet the expectation.

Past experiences

As discussed in an earlier article evaluation and budgeting can be linked in different ways and in different stages of the budget process. Evaluations can be used to point at needs, to provide information on proposals, to analyze processes and to get information of the results of the budget spending (Bastoe, 1999). In a World Bank paper on the connection between evaluation and budgeting, Marc Robinson concludes that "unfortunately, the potential value of evaluation as a budgeting tool has not been realized in practice. In part, this is because evaluation has often not been sufficiently tailored to the needs of budget decision makers" (IEG, 2014). Our experience suggests that evaluations in development cooperation mostly serve to facilitate improvement *within* ongoing or future aid interventions rather than to guide the allocation *between* different interventions.

A review of the Norwegian Government's 2015 budget proposal for the Ministry of Foreign Affairs demonstrates the limited relevance of evaluations in budget allocations. The almost 200 page document mentions evaluations and evaluation findings 31 times, mostly in general terms. Encouraging, 10 of the references concern intentions to increase the use of evaluations. References to specific evaluation findings mostly focus on the positive findings, apparently to legitimize rather than to guide budget allocations.

What would it take to meet the expectation?

In our efforts to facilitate better use of evaluations, we believe that more importance should be attached to decisions made very early in the evaluation process. Attention is usually primarily given to what happens between developing the ToR and the presentation of the final report. This is natural as it is the most labor-intensive phase of an evaluation. However, based on experiences, we believe that relevance and utilization of

1. Political platform for a government formed by the Conservative Party and the Progress Party, Sundvolden, 7 October 2013. Available at https://www.regjeringen.no/no/dokumenter/politisk-plattform/id743014.

evaluations are to large degree determined by strategic choices made well before even starting to develop a ToR. Careful consideration of what, when and how to evaluate may be crucial for the effective use of evaluations. This should be based on good knowledge of both policy and budget processes and in close dialogue with the relevant units - not necessarily to reach consensus, but rather to explore knowledge gaps related to budgeting.

What and when to evaluate

As a starting point, it is also important to understand cultural aspect and "beliefs" among decision-makers. There are institutions and interventions, for which funding is determined by many factors more or less unaffected by evaluation results – e.g. political champions, foreign policy interests, or simply that the institutions and interventions in question are considered the only one 'available' for the government if it wants to pursue certain objectives. They are like "sacred cows", almost untouchable. In such cases evaluations are probably not likely to make a difference, at least not in terms of budget allocations. In Norwegian development cooperation, due to historical reasons and political ties, this seems to be the case for certain UN organizations and certain NGOs. In broad terms, the allocation of aid funds for some organizations seems to be unaffected by their performance and ability to achieve results. To meet the government's evaluation expectations of evaluations with budgetary consequences, one should perhaps not even attempt to evaluate these institutions and interventions. Alternatively, one should evaluate for other purposes than influencing budgets.

The room for evaluations to make a difference may also depend a lot on timing. At the most basic level, in a program evaluation, the best time may be well before decision makers start the discussions about prolonging into a new multi-year agreement. This is an argument against the typical ex-post evaluation, often happening at the end of, or just after a program period, where an informal decision on continuation or not are already made – however, ex-post evaluations are key to accountability for results, which again may be a key dimension in some budgetary processes. Mid-term evaluations might perhaps be a better timing when it comes to relevance for budget allocations. One challenge in this regard is that the cycle of budgetary processes – or at least the window of opportunity for evaluations to make a difference – may be well shorter than typical evaluations.

Timing is also relevant in broader terms, as shifting political interests may influence the possibility for evaluations to make a difference. Two recent examples are the evaluations of the Norwegian private investment fund (Norfund) and of the support to primary education through Unicef and the Global Partnership for Education (GPE). They are all among the largest recipients of Norwegian aid, and already among the 'sacred cows' mentioned above. Both evaluations have expressed doubts about the relevance and/or effectiveness of these organizations. However, there is no indication that the evaluations will affect the budget allocations – at least not in the short time perspective. This may perhaps be explained not only by key staff in the Ministry having close affiliations and strong confidence in these organizations, but also that trade and investment, and primary education for girls are among the two highest priorities for the current government, and there is simply few other alternatives available for the government to channel large funds towards these objectives.

An example of good timing may be our recent Myanmar baseline study, emphasizing the critical lack of contextual analysis in Norwegian aid to Myanmar. This is one of the evaluations that have sparked most interest with many positive spinoffs and possible effects on future budget allocations. Given the relatively low costs, it has given a lot of value for money. The explanation may be the timing. Norway has recently engaged heavily in Myanmar and key actors seem very eager to do things right, and as the programme is still in its very early stage it is not too late to make changes.

Tentative conclusions

Based on this brief reflection on experiences related to the relevance of evaluation for budget allocations we claim that it is important to pay attention on the early decisions about what and when to evaluate. Those decisions are worth more attention than today, where one tend to focus more on issues relating to decisions made later in the process, such as the ToR, the organization of the evaluation, the choice of evaluators etc. This also implies making choices about what *not* to evaluate – and certainly not to evaluate anything only because the program cycle suggests so and stakeholders expect it. It also involves having to accept that in some cases, forces are too 'strong' for evaluations to make a difference anyway.

References

Bastoe, P. O. (1999). *Linking Evaluation with Strategic Planning, Budgeting, Monitoring and Auditing*. Chapter in Richard Boyle and Donald Lemaire (eds). Building Effective Evaluation Capacity. Transaction Publishers.

Robinson, M. (2014). *Connecting Evaluation and Budgeting*. ECD Working Paper. Independent Evaluation Group. World Bank.

THE POTENTIAL ROLE OF CONSEQUENTIAL VALIDITY IN EVALUATION

Sebastian Lemire

Validity is a perennial topic in evaluation. Most recently, Chen, Donaldson and Mark (2011) co-edited a special edition on "validity in outcome evaluation – advancing new ideas in both theory and practice of validity investigations". In their respective contributions, Stewart Donaldson and Jennifer Greene advocated increased attention to evaluation use in validity investigations. Responding to this call, this brief article proposes a stronger role for consequential validity in evaluation.

What is consequential validity?

Consequential validity is nothing new. Messick (1995) originally introduced consequences to the validity argument in the context of educational assessment, stating that *"the consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (p. 6)*. Underlying Messick's interest in consequential validity is a distinction between the "evidential" (psychometric) and "consequential" (value-driven) aspects of test validity (1994) – both of which resonate with fundamental features of the evaluation discipline.

Despite the important role of consequential validity in the context of teaching assessment, the concept has not gained noticeable traction in the broader context of evaluation. To be sure, evaluators consider the consequences of their evaluation practice, especially in the context of conflict prevention and peace building activities ('do no harm') as well as in developmental evaluation and participatory evaluation more generally (Patton, 2011).

However, in regular evaluation practice, considerations of consequences are rarely partnered – at least explicitly – with validity considerations. The evidential and consequential are kept separate. This is unfortunate because reflecting upon this connection would heighten awareness of how and in what ways certain evaluation designs and procedures tend to promote positive effects and/or impede adverse consequences.

Painting in broad strokes, two central aspects of the evaluation landscape collectively call for increased attention to consequential validity in evaluation: the commitment to use and the persistent focus on methodological rigor. Consider the persistent role of 'systematic method' in evaluation practice: whereas beliefs about what constitutes methodological rigor differ, most evaluators commit to the idea that sound methodology is front and center among evaluation quality criteria, especially from a summative standpoint.

However another defining feature of evaluation is formative and addresses evaluation use. The numerous conferences, articles and books dedicated to utilization concepts and issues speak to this point. Evaluation is an applied field that situates actual use of evaluations at the very core of its practice. Indeed, many consider the active use of evaluations in making decisions about projects, programs, and policies intrinsic to the quality of evaluation. Others warn that a dominant preoccupation with use may undermine impartiality.

Thus, despite their central roles, methodological rigor and evaluation use are in practice often considered in parallel or even in conflict, rather than constituting an integrative whole. The metaphor of independent method and use branches on a shared evaluation theory tree serves well to illustrate their intertwined, yet separated purposes and development in evaluation theory (Christie and Alkin, 2013). This separation is unfortunate because the two are in so many ways intertwined and interdependent.

Taken collectively, then, both methodological rigor and evaluation use (however defined) comprise key features of evaluation practice. From this perspective, the concept of validity implies a similar commitment. Inspired, then, by Messick, we define consequential validity in the context of evaluation, as: *the extent to which positive consequences result (and adverse effects minimized) from evaluation design and implementation*.

Formulated this way, consequential validity unites the evidential (methodological rigor) and consequential aspects (use) of evaluation process and practice by focusing attention on how positive results or adverse consequences of evaluative information can be traced back to the design and implementation of the evaluation. By bridging the evidential and the consequential, consequential validity connects the rigor and the use of evaluation. The benefit of doing so is a better understanding of how and why different aspects of our practice result in adverse consequences.

Examining consequential validity in evaluation

For consequential validity to even get off the ground some boundary probing is called for. Specifically the following defining dimensions of consequential validity are worth consideration:

- *Level.* Does consequential validity pertain to individual evaluations, approaches to evaluation (e.g. theory-based evaluation) or to the field of evaluation in general?
- *Scope.* Can consequential validity measured by the consequences of the evaluation be disentangled from the characteristics of the evaluand or the behavior of the evaluee?
- *Type of consequences.* How much weight should be ascribed to negative and positive, intended and unintended consequences?

Based on the injunction of the Hippocratic Oath (First do no harm) and since the 'doing good' principle raises highly speculative and controversial ethical questions a focus on the potentially adverse effects of evaluations should be privileged. In doing so, the following deceptively simple questions are relevant:

- What are the most salient adverse consequences of the evaluation?

- How are these adverse consequences connected to the questions and design features used to structured the evaluation?
- Have the risks of adverse consequences been minimized using procedures that are consistent with sound research design?
- Are the associated risks reasonable in relation to anticipated benefits?
- Are safeguards in place to protect the public from potentially adverse effects?
- Has the appropriate type of use or decision-making supported by the evaluation been specified.

Taken collectively, these questions promote consideration of how and in what way adverse consequences tie back to limitations or flaws in evaluation design, process or practice. Such flaws might have to do with framing the evaluation, designing it, the choice of analytical strategy, or the focus on the types of decisions informed by the evaluation.

What are the implications?

Adapting consequential validity in evaluation raises two concerns: (1) are we diluting the concept of validity by introducing yet another type of validity and (2) are we overburdening the evaluation practitioner by demanding this type of validity investigation (Shepard, 1997).

The first implication is theoretical and motivated by the already broad and varied landscape of validity. However, the concept of consequential validity is already well established in educational measurement without creating confusion and pursuing consequential validity is more of a refinement than an expansion of the concept of validity in evaluation.

The second implication carries more weight. Unlike educational measurement and assessment, in which the risks are more readily identifiable and quantifiable, adverse consequences associated with other types of evaluations may be more ambiguous and less predictable. As a result, it might be too burdensome to examine consequential validity as part of often tightly budgeted evaluations.

While it may be foolhardy to suggest that evaluators should speculate about all possible adverse consequences of their work, curtailing its most salient adverse consequences may still be feasible. The benefit of doing so would not only serve to enhance the credibility of specific evaluation studies, but also further the field of evaluation in general.

References

Chen, H.-T., Donaldson, S., and Melvin, M. (2011). *Validity in outcome evaluation, advancing new ideas in both theory and practice of validity investigations* (p. 130). New Directions for Evaluation.

Christie, C. and Alkin, M. (2013). *An evaluation theory tree*. In M.C. Alkin (Ed.) Evaluation Roots – A Wider Perspective of Theorists' Views and Influences. Thousand Oaks, CA: Sage.

Messick, S. (1994). *The interplay of evidence and consequences in the validation of performance assessments*. Educational Researcher, 23(2), pp. 13–23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(2), pp. 5–8.

Patton, M. Q. (2011). *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovations*. New York, NY: The Guilford Press.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), pp. 4–13.

EVALUATION LOGICS IN THE THIRD SECTOR[1]

Matthew Hall

Third sector organizations (TSOs) are neither part of government nor for-profit businesses. Although diverse, they typically conduct activities geared towards a social purpose, e.g. provision of welfare services, international development programs, human rights advocacy, etc. They also often address the needs of groups that are marginalized or neglected by government and business. Evaluations of third sector organizations are used to promote their work; demonstrate accountability to donors and beneficiaries and assess performance.

Third sector debates over the merits and drawbacks of specific evaluation techniques tend to focus on technical aspects, e.g. the merits of quantitative indicators vs. case studies, or the appropriateness of randomized controlled trials. There has been less debate, however, about the normative beliefs underpinning evaluation practices in TSOs (Bouchard, 2009a; 2009b; Eme, 2009).

This article identifies the ideals embedded in different evaluation approaches so as to develop a typology for evaluations in the third sector. Three 'ideal-type' logics dominate: scientific, bureaucratic and learning. Together they capture a wide range of evalu-

1 This article is an updated and condensed version of Hall (2014). Evaluation logics in the third sector. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 25, pp. 307–336.

ation methods, such as the logical framework (e.g., Rosenberg & Posner, 1979); the most significant change stories (e.g., Dart & Davies, 2005), the social return on investment approach or SROI (e.g., New Economics Foundation, 2007); scorecards (e.g., Kaplan, 2001); outcome frameworks (e.g., Urban Institute, 2006); sundry participatory methods (e.g., Keystone Accountability, undated); and the best available charitable option or BCA (e.g., Acumen Fund, 2007).

The first category (*scientific evaluation logic*) focuses on systematic observation, the gathering of observable and measurable evidence, and a concern with objective and robust experimental procedures. Its evaluation ideals include those of proof, objectivity and conflict reduction. The evaluator's role is akin to that of a scientist. For example, such evaluation techniques as the social return on investment focus on attributing intended outcomes to third sector projects, through experimental methods. Such methods are not well adapted to TSOs as they typically have limited financial resources to devote to elaborate techniques and the extensive data collection and analysis that they require. Donors are reluctant to provide funds for such spending often classified as overhead or 'administrative' expenditure. Furthermore TSOs have limited capacity in research methods such as econometrics and statistics.

The second category (*bureaucratic evaluation logic*) is focused on careful observation of complex, step-by-step procedures, limiting of deviations from such procedures, and analysis of the achievement of intended objectives. Its evaluation ideals are those of "sequentiality", the primacy of intended effects and hierarchy. Under this model, the evaluator's role is that of an examiner. For example, a wide range of evaluation techniques, including Logical Framework Analysis (LFA), the Balanced Scorecard, the Common Outcome Framework (Urban Institute, 2006), and the Best Available Charitable Option (Acumen Fund, 2007) all focus on fitting the evaluation of projects into categories predefined by the evaluation method itself (such as the LFA, which maps projects on the categories of inputs, outputs, purposes and goals). This approach can be suitable for simple and well-defined projects undertaken by TSOs (such as a vaccination program), but can be problematic for complex and long-term projects with unpredictable outcomes, such as efforts to change government policy or long-term social development projects in rural communities.

The third category (*learning evaluation logic*) privileges openness to change and the unexpected, the incorporation and consideration of a wide range of views and perspectives, and a focus on lay rather than professional expertise. Its evaluation ideals are those of richness, belief revision, and egalitarianism, and the evaluator's role is primarily that of a facilitator. For example, techniques such as the Most Significant Change and the Impact Planning, Assessment and Learning method (Keystone, undated: 3) focus on both the intended and unintended, and positive and negative, outcomes that can emerge during a project.

The learning evaluation logic is especially well attuned to the typical characteristics of TSOs, particularly the focus on unexpected outcomes and the fewer resource and expertise requirements. However, the evaluative outputs of these techniques can fall short of donors' expectations given the low public tolerance for negative outcomes and their desire for rigorous verification of tangible outcomes.

Developing an understanding of evaluation logics in the third sector is critically important given the diverse value frameworks of TSO stakeholders. Only where these differences are confronted and resolved can principled agreement be reached on the choice of evaluation techniques and the capacity building implications for TSOs that lack resources and expertise in evaluation. Identifying the normative properties of different evaluation approaches can help to unpack debates between donors, TSOs, governments, clients and other stakeholders over the merits of particular evaluation approaches.

In particular, it can help distinguish between debates centered on differences in evaluation ideals (e.g., a belief in the outright superiority of particular methods or forms of data) and those that are more methodological in nature (e.g., disagreements about the validity with which specific data or methods are used or applied). The ability to differentiate between ideological and methodological criticisms helps explain the distinct rationales and viewpoints advanced by stakeholders in TSOs.

For instance, an ideological critique, such as a demand for 'proof' of outcomes from donors is likely to prove intractable even where TSOs seek to be responsive. In contrast, a methodological critique, such as a desire for more transparency about how a method was used or applied, could be addressed by changing the evaluation technique to improve its validity in the eyes of important stakeholders such as donors.

A focus on evaluation logics also directs attention to how different evaluation ideals can privilege different kinds of knowledge and the desired process for knowledge generation during evaluation processes. This is critical because evaluations influence the legitimacy of TSOs in the eyes of different stakeholders. For example, stakeholders can demand quantitative information demonstrating the impact of TSOs, which can conflict with techniques focused on story-telling and dialogue. Understanding the evaluation logics underpinning these preferences can potentially help TSO stakeholders to negotiate and reconcile differences by balancing or blending different types of information and methods of knowledge generation.

Such analysis also directs explicit attention to the conceptions of expertise and resource requirements accompanying different evaluation logics. This is clearly appropriate at a practical level where financial and other resources are typically limited for TSOs. Different conceptions of expertise and 'valid' information can also elevate the interests of certain stakeholders in TSOs whilst disenfranchising others, particularly those marginalized stakeholders the TSO is focused upon.

This may occur where complex evaluation techniques full of jargon and arcane concepts are used. While they may be readily familiar to evaluators and donors they are difficult to get across the diverse contexts, languages and cultures within which TSOs operate (Wallace et al., 2007). As such, greater understanding of the operating context and its implications for the level and types of expertise implied by particular evaluation logics is important for evaluation to serve those whose interests are the most legitimate.

References

Acumen Fund (2007). *The Best Available Charitable Option.* Accessed 19 September 2011, from http://www.acumenfund.org/uploads/assets/documents/BACO%20Concept%20Paper%20final_B1cNOVEM.pdf.

Bouchard, J. M. (2009a). The worth of the social economy. In Bouchard, J. M. *The Worth of the Social Economy: An International Perspective* (pp. 11–18). P.I.E Peter Lang, Brussels.

Bouchard, J. M. (2009b). The evaluation of the social economy in Quebec, with regards to stakeholders, mission and organizational identity. In Bouchard, J. M. *The Worth of the Social Economy: An International Perspective* (pp. 111–132). P.I.E Peter Lang, Brussels.

Dart, J., & Davies, R. (2003). A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique, *American Journal of Evaluation*, 24, pp. 137–155.

Eme, B. (2009). Miseries and worth of the evaluation of the social and solidarity-based economy: for a paradigm of communicational evaluation. In Bouchard, J. M. *The Worth of the Social Economy: An International Perspective* (pp. 63–86). P.I.E Peter Lang, Brussels.

Hall, M. (2014). Evaluation logics in the third sector. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 25, pp. 307–336.

Kaplan, R. S. (2001). Strategic performance measurement and management in third sector organizations, *Nonprofit Management & Leadership*, 11(3), pp. 353–371.

Keystone (undated). *Impact Planning, Assessment and Learning.* Accessed 19 September 2011, from http://www.keystoneaccountability.org/sites/default/files/1%20IPAL%20overview%20and%20service%20offering_0.pdf.

New Economics Foundation (NEF) (2007). *Measuring Real Value: a DIY Guide to Social Return on Investment.*

Roberts Enterprise Development Fund (REDF) (2001). *Social Return on Investment Methodology: Analyzing the Value of Social Purpose Enterprise Within a Social Return on Investment Framework.*

Rosenberg, L. J. and Posner, L. D. (1979). *The Logical Framework: A Manager's Guide to a Scientific Approach to Design and Evaluation.*

Urban Institute (2006). *Building a Common Outcome Framework to Measure Nonprofit Performance.* Accessed 19 September 2011, from http://www.urban.org/UploadedPDF/411404_Nonprofit_Performance.pdf,

NEWS

During the AGM, held 20th November 2015, new membership fees were approved. From January 1, members will have the option to chose between the hard copy of Evaluation Journal and between the e-version, where reduced fees will apply. The overview of the new membership fees is presented in the table below.

Current membership fee structure		New membership fee structure	
Type of membership	**Fee (EUR)**	**Type of membership**	**Fee (EUR)**
1 year-individual (print copy + electronic access)	150	1 year-individual (print copy + electronic access)	150
–	–	1 year-individual – electronic-only	140
2 year-individual (print copy + electronic access)	250	2 year-individual (print copy + electronic access)	250
–	–	2 year-individual – electronic-only	220
1 year-student	62.50	1 year-student – electronic-only	62.50
1 year-institutional	1,200	1 year-institutional – print copy	1,200

During the Annual General Meeting, the election results were announced.

The AGM also provided an opportunity to extend special thanks and appreciation to departing Board members – Barbara Befani (Secretary General), Kim Forss (Treasurer) and Liisa Horelli – for their exceptional commitment and substantive contributions to the Society. Election results for board member positions vacant as of January 1st 2016 were also announced. The new elected board members are: Bastiaan de Laat (Vice President), Julia Brummer, Laura Tagle and Jos Vaessen. (see bios below).

Julia Bruemmer

Julia has some eight years of experience in evaluation-related work, in the field of international development. She worked for the evaluation section of the ILO's International Programme on the Elimination of Child Labour and currently holds the position of Project Monitoring Officer at the Lutheran World Federation. Next to her regular job, she takes on independent consultancy work, mostly providing M&E trainings and conducting independent evaluations. Julia moreover is a PhD fellow in "Governance and Policy Analysis" at the Maastricht Graduate School of Governance. Her research focuses on impact evaluation methods. She has previously been involved in the EES Board as a co-opted board member. Since 2014 she leads the EES Thematic Working Group for Emerging Evaluators.

Laura Tagle

Laura has been involved in the evaluation field for over 20 years. She has studied and worked at both the national (Italy) and at the international level. Her work focuses on the responsibility of the public sector in evaluation and on the evaluation of regional development. She is currently developing an approach to evaluating public policies from the point of view of local actors. She served as a board member at the Italian Evaliation Asociation (AIV) and at the International Development Evaluation Association (IDEAS). She co-founded the LVD (Laboratorio di Valutazione Democratica), where she engages in exploring of the issues involved in democratic evaluation practices.

Jozef "Jos" Vaessen

Jos Vaessen (Ph.D. Maastricht University) is Principal Evaluation Specialist at the Internal Oversight Service of UNESCO in Paris and lecturer at Maastricht University, The Netherlands. After completing his M.Sc. in 1997 (Wageningen University) and prior to starting his current position at UNESCO in 2011, he has been involved in research, teaching and evaluation activities in the field of international development at Antwerp University and, more recently, Maastricht University. Over the last fifteen years or so, he has worked for several multilateral and bilateral international organizations mostly on evaluation-related assignments. His fields of interest include: theory and practice evaluation, impact evaluation, rural development and environment. In addition to managing and conducting evaluations Jos regularly serves on reference groups of evaluations of different organizations. He has been (co-) author of more than 30 publications, including three books. Recent publications include: Impact evaluations and development – NONIE guidance on impact evaluation (2009, co-author, with F. Leeuw; NONIE), Mind the gap: perspectives on policy evaluation and the social sciences (2009, co-editor,with F. Leeuw; Transaction Publishers), Dealing with complexity in development evaluation: a practical approach (2015, co-editor with M. Bamberger and E. Raimondo; SAGE Publications).

Bastiaan de Laat

Bastiaan de Laat (PhD) is Evaluation Expert and Team Leader at the European Investment Bank (EIB) where over the past couple of years he has been in charge of major evaluations in important areas such as Climate Action, SME support and Technical Assistance. He has a longstanding experience in evaluation as well as in foresight. Founder-director of the French subsidiary of the Technopolis Group (1998–2006) he led many evaluations for and provided policy advice to a great variety of local, national and international public bodies. He trained several hundreds of European Commission staff and national government officials in evaluation and designed monitoring and evaluation systems for various public organisations. Before joining the EIB he worked as Evaluator at the Council of Europe Development Bank. He has developed tools and performed programme, policy and regulatory evaluations, both ex ante and ex post, in a variety of fields. He has also made several academic contributions, most recently with articles on evaluation use and on the "Tricky Triangle", on the relationships between evaluator, evaluation commissioner and evaluand.
Bastiaan served as EES Secretary General 2011–14, was programme coordinator of the 2010 EES Conference in Prague, and general coordinator of the 2012 (Helsinki) and 2014 (Dublin) EES Conferences.

GUIDANCE
TO CONTRIBUTORS

Brevity and nimbleness are the hallmarks of Connections: published articles are normally 800–1,200 words long. Even shorter contributions (news items; opinion pieces; book reviews and letters to the editor) are accepted with a view to promote debate and connect evaluators within Europe and beyond. While Connections is not a peer reviewed publication only articles that add to knowledge about the theory, methods and practices of evaluation should be submitted.

Contributions that highlight European values and evaluation practices are given priority but Connections also reaches out beyond Europe to the international evaluation community and favours articles reflecting a diversity of perspectives. Within the limits of copyrights agreements articles that summarize in a cogent way the substantive content of published (or to be published) studies, papers, book chapters, etc. are welcome (with suitable attribution).

Individuals or organizations wishing to sponsor special issues about an evaluation theme or topic of contemporary interest should contact the EES Secretariat (secretariat@europeanevaluation.org). Such special issues usually consist of 6–8 articles in addition to a guest editorial. A Presidential letter may be included. The guest editor(s) are responsible for the quality of the material and the timeliness of submissions. The regular editorial team ensures that special issues meet 'Connections' standards and takes care of copy editing.

To facilitate copy editing, authors are encouraged to use end notes rather than footnotes and to use the APA style guide for references. Here are some examples:

For books: Bergmann, I. (1997). Attention deficit disorder. In *The new Encyclopedia Britannica* (Vol. 26, pp. 501–508). Chicago, IL: Encyclopedia Britannica.

For journal articles: Rindermann, H., & Ceci, S. J. (2009). Educational policy and country outcomes in international cognitive competence studies. *Perspectives on Psychological Science*, 4(6), 551–568. doi:10.111/j1745-6924. 2009.01165.x

Website: About.com Islam, (2014). Evils of Gossip and Backbiting in Islam. Retrieved 12 June 2014, from http://islam.about.com/od/familycommunity/a/Gossip-Backbiting.htm.

www.ees2016.eu

12th EES Biennial Conference

Evaluation Futures in Europe and beyond
Connectivity, Innovation and Use

MECC Maastricht, the Netherlands
26–30 September, 2016

EES EUROPEAN EVALUATION SOCIETY

THE EES 12th BIENNIAL CONFERENCE

The EES 12th Biennial Conference **"Evaluation Futures in Europe and Beyond: Connectivity, Innovation and Use"** promises to be the evaluation event of 2016. It will be held in Maastricht, The Netherlands, in the same country where the first EES conference was held over 20 years ago. It will take stock of the current state of affairs in evaluation politics, capacity, systems, research, methods, communication and use; as well as sketch future directions.

The overall programme is divided into 4 thematic strands, with two cross-cutting themes (Europe and Gender Equality):

- **Strand One ("Evaluation Ethics, Governance, and Professionalism")** addresses "the rules of the game": high-level standards and normative frameworks for evaluators as well as commissioners and policy makers.
- **Strand Two ("Evaluation Systems, Organisations and Partnerships")** relates to the institutional architecture of evaluations, for example the development of evaluation systems or multi-partner and complex network configurations.
- **Strand Three ("Evaluation Methods and Research")** reflects on innovative ways to design and conduct evaluations, as well as on evaluation theories and approaches based on academic, discipline-based traditions.
- **Strand Four ("Evaluation Use, Communication and Outreach")** highlights experiences showing that evaluations can have an impact, not just on policy making but also on community empowerment together with other desirable (or undesirable) outcomes.

Our **keynote and plenary session** speakers will take us on a memorable journey of entertaining enlightenment (or enlightening entertainment):

- **Hans Rosling**, the man who makes "statistics sing" or "data dance" will teach us a thing or two about communicating evaluation findings. But most importantly he will make us realize how easy it is to be wrong by uncovering ignorance and biases affecting everyone, even "the educated".
- **Elliot Stern** will represent "evaluation" in a plenary panel chaired by **Frans Leeuw**, joining two high-level representatives of other social sciences who will compete, or perhaps cooperate, among themselves and with "evaluation" to provide the most appropriate and highest quality advice to a fictitious prime minister about to design a new policy.
- Our **EU keynote speaker** will remind us what is at stake for Europe: the value Evaluation holds for Europe-wide cohesion, democracy and accountability.
- **Claire Hutchings** will pull the threads of the conference together and draw on her experience as an NGO commissioner, a methodological innovator, and an M&E community developer, to tell us what it takes for innovation in evaluation to spread, and for evaluations to be pervasive and persuasive, making a difference "on the ground".

We look forward to welcoming you in Maastricht!

THE AUTHORS

Ole Winckler Andersen

is Deputy Permanent Representative at the Danish Delegation to the OECD. He holds a M.Sc. in Economics and a Ph.D. in Public Administration and has worked for the Danish Ministry of Foreign Affairs for the last 20 years, including as Head of Evaluation Department (2007 – 13). Before that he was assistant and associate professor at Roskilde University in Denmark. He has managed a number of evaluations and has served as team leader for missions to Asian, African and Latin American countries. He has also been member of several management committees for international development evaluations. He has published on evaluation in various international journals and recently co-edited Evaluation Methodologies for Aid in Conflict (Routledge 2014).

Dr. Chris Barnett

is the Director of the Centre for Development Impact (CDI), a joint initiative between the Institute of Development Studies, ITAD and the University of East Anglia. CDI focuses on improving the evaluation and understanding of impact, including agenda setting and spearheading learning and innovation on methodological debates. Chris is also Technical Director at ITAD. Chris has had extensive experience of leading and managing evaluations – especially in Africa – and is currently the Project Director for the quasi-experimental impact evaluation of the Millennium Villages Project (Ghana, 2012–2017); the case-based evaluation of a governance fund (Malawi, 2012–2016); and the developmental evaluation of the M4D governance programme (Nigeria, 2014–2018). He also works on the monitoring, verification and evaluation of Ideas to Impact, a five-year programme of innovation prizes (Global, 2013–2018).

Per Øyvind Bastoe

is the Evaluation Director at Norad with responsibility to evaluate all aspects of Norwegian development policy. He has broad experience from international development and evaluation and has previously held senior positions in other parts of the Norwegian government administration, in the World Bank and the Asian Development Bank. Before taking up his current position, he served at the Executive Board of the Inter-American Development Bank and the Inter-American Investment Cooperation on behalf of a group of European countries. He is a member of the International Evaluation Research Group and has published several books and articles on development policy, evaluation and organizational change. His most recent book "New Challenges and New Roles – Development Financing in the 21st Century" discusses the rapidly changing landscape of development financing and the possibilities and limitations for the Nordic countries.

Agathe Devaux

Agathe Devaux-Spatarakis is the Scientific manager of impact evaluations of agronomic research for development in Cirad. She completed in 2014 a PhD in political science. Her research focus was on the unfolding of RCTs in France and the evidence-based policy movement within the French government. While conducting her PhD, she was an evaluation consultant for the consulting firm Eureval for 4 years and participated to a dozen evaluations of public policies for local, regional, national level of government and the European Commission. She is also a lecturer in the institute of political science in Lyon for the master degree in evaluation of public policies.

Øyvind Eggen

Øyvind Eggen is policy director for evaluation at the Norwegian agency for development cooperation (Norad). Prior to his current position he was senior researcher at Norwegian Institute of International Affairs, specializing in Western aid discourse and policy. He holds a PhD in development studies, addressing the effects of aid on governance in Malawi.

Matthew Hall

is Associate Professor in Accounting at the London School of Economics and Political Science, where he has worked since 2006. Matthew's research interests relate to management accounting, performance measurement and the use of accounting information in public policy debates. His current research is focused primarily on the development and use of performance measurement in the third sector through in-depth case studies in Australia, the UK and the US. In particular, Matthew is examining the development of techniques designed to measure social value (such as social return on investment) and how they become implicated in the operations of NGOs, impact assessment and discussions of NGO effectiveness more broadly.

Sebastian Lemire

is a doctoral candidate at the University of California, Los Angeles. His area of interest revolves around theory based evaluation, research synthesis, and evaluation capacity building. Sebastian has published on these topics in the American Journal of Evaluation, Evaluation, and the Canadian Journal of Program Evaluation.

Riitta Oksanen,

EES Vice President, is a senior advisor on development evaluation in the Ministry for Foreign Affairs of Finland. She previously held the post of director for development policy in the Ministry and recently chaired the OECD/DAC Evaluation Network task team on evaluation capacity development. She has also acted as advisor on management and effectiveness of development cooperation; served in Finland's permanent EU delegation as counsellor responsible for EU development policy and cooperation, and chaired the Council's working group on development cooperation during the Finnish EU Presidency in 2006. Before joining the Ministry in 1999 she worked as a consultant specialising in development cooperation planning, management and evaluation. She is a University of Helsinki graduate specialized in marketing, business administration and economics with emphasis on the forestry sector.

Dr. Anna Paterson

is a political scientist and freelance evaluator and researcher with extensive experience of a range of evaluation methods and types especially for governance and peacebuilding interventions. She has worked in East and West Africa, Central Asia, Afghanistan and Pakistan. Anna worked for two years as a field researcher in Afghanistan Research and Evaluation Unit (AREU) and subsequently conducted her PhD research in Afghanistan. She then worked for DFID, for two years as a Research Evidence Broker and for a further year as an Evaluation Adviser in DFID Nigeria, before becoming a consultant.

Robert ('Bob') Picciotto,

(UK) Professor, Kings College (London) was Director General of the World Bank's Independent Evaluation Group from 1992 to 2002. He previously served as Vice President, Corporate Planning and Budgeting and Director, Projects in three of the World Bank's Regions. He currently sits on the United Kingdom Evaluation Society Council and the European Evaluation Society's board. He serves as senior evaluation adviser to the International Fund for Agricultural Development and the Global Environment Fund. He is also a member of the International Advisory Committee on Development Impact which reports to the Secretary of State for International Development of the United Kingdom.