Article

Evaluation
17(3) 261–276

DOI: 10.1177/1356389011410518
evi.sagepub.com
SAGE

Evaluating aid: An adolescent domain of practice

Livingston Armytage
Centre for Judicial Studies, Australia

Abstract
This article addresses the question '*Does aid work?*' by asking '*How do we know if it works?*' Despite substantial refinement in evaluation approaches, evaluation remains without any orthodoxy about how to assess effectiveness. The article examines the purposes of evaluation to discern unresolved tensions between accountability and an organizational learning approach. This is framed by the current vogue of Managing for Development Results, integral to the Paris Declaration on Aid Effectiveness in attaining the Millennium Development Goals. The article appraises the 'paradigm war' revealed in debates over methodology and the utility of the logical framework. It critiques limitations in development evaluation doctrine to highlight profound uncertainties of attribution and causality. The 'evaluation gap' between the rhetoric of donors and their evaluation practice, unsurprisingly, inhibits the evaluation of aid effectiveness.

Keywords
development aid, development effectiveness, governance law and justice, monitoring and evaluation, overseas development assistance

Introduction

This article addresses the question '*Does aid work?*' by analysing the theory and practice of development evaluation to ask '*How do we know if it works?*'

Over the past decade there has been substantial expansion and refinement in the evaluation of international development. At the same time, evaluation remains in a transitional state without any orthodoxy as to how to assess effectiveness. In this article I consider the purposes of evaluation with particular regard to governance and justice reforms in order to clarify unresolved tensions between the accountability (audit) approach and emerging effectiveness (learning) approaches. This is framed by the current vogue for Managing for Development Results which is integral to the Paris Declaration on Aid Effectiveness and the attainment of the Millennium Development Goals. I appraise the 'paradigm war' between positivists and constructivists revealed in simmering debates

Corresponding author:
Livingston Armytage, Director of the Centre for Judicial Studies, Sydney, Australia.
Email: larmytage@yahoo.com

over methodology and the utility of the logical framework approach and critique limitations in development evaluation doctrine that concentrate on serving new public management principles. Of arguably most concern is the risk that recent emphases on monitoring are overshadowing the more vexing challenges of evaluation.

I conclude by finding a marked 'evaluation gap' between the new rhetorical posture of the international community and the state of its evaluation practice, with ensuing confusion over fundamental aspects of purpose, approach and methodology. Unsurprisingly, this confusion inhibits the evaluation of aid effectiveness.

Evaluation and monitoring

Evaluation provides the means for accountability as well as improvement in development assistance. It therefore requires clear and consistent answers to the questions '*what'* reform is supposed to achieve, and '*how*' success is to be measured. Evaluation addresses the question: '*Does aid work?'*

For some, evaluation is primarily about information for decision making. Thus according to the Organisation of Economic Cooperation and Development:

> *Evaluation* is the systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation and results. The aim is to determine the relevance and fulfilment of objectives, development efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors. (OECD-DAC, 2002: 2)

However, evaluation is also inescapably a normative process, which depends on a framework of values. Evaluation requires normative criteria to make judgments about activities and their value (Armytage, 1996: 186). It goes beyond gathering data to assess quality, value, worth and effectiveness (United Nations Evaluation Group, 2005).

In development practice, evaluation is frequently associated with monitoring. While these activities are often contiguous, they are separate. *Monitoring* is the process of observing and reporting on something over a period of time. It is defined as:

> A continuing function that uses systematic collection of data on specified indicators to provide management and the main stakeholders of an ongoing development intervention with indications of the extent of progress and achievement of objectives and progress in the use of allocated funds. (OECD-DAC, 2002: 4)

It is generally accepted that monitoring focuses on enabling managers to track what is happening with their projects and to check on progress towards the achievement of objectives. Evaluation is a more purposeful activity which is concerned with learning lessons from past experience for the future and providing accountability (Cracknell, 2000).

It is important to maintain a clear distinction between *monitoring*, which is mainly concerned with tracking efficient delivery of 'inputs' and 'outputs', and *evaluation* which is mainly concerned with assessing the worth of 'outcomes' and 'impacts'. Evaluation is a contextualized, purposeful, normative function, which usually builds on data supplied through monitoring. Donors need to both monitor the performance of their contractors and evaluate the performance of their assistance. The former has an audit or accountability purpose, while the latter has a learning or effectiveness purpose. While both may co-exist, their distinction is often muddied in practice.

Purpose and models

Evaluation serves a variety of purposes (Leeuw and Furubo, 2008). In the present development context, the most important of these are accountability and effectiveness.

The OECD-DAC *Principles for Evaluation of Development Assistance* based on a consensus of the world's major development agencies specified the purposes for evaluation as being to improve future aid policy and programmes through feedback of lessons learned; and to provide a basis for accountability, including the provision of information to the public (OECD-DAC, 1991). Picciotto similarly argued in reviewing the World Bank's experience of evaluation:

> In a nutshell, feedback and follow-up were central to the conception of (the Office of Evaluation and Development) OED from day one. Feedback is about learning and follow-up is about accountability. They are two sides of the same coin . . . (Picciotto, 2002a: 1)

These purposes are not necessarily points on a continuum. Rather, the evidence of practice demonstrates that they are distinct and at odds. I will argue, as others have done (Binnendijk, 2000: 20), that the major purposes for audit and accountability on the one hand, and learning and improvement on the other, may not be compatible because they operationalize the evaluation function using different data, methodologies and incentives.

The divergent purposes of evaluation rooted in different models of practice (Norton, 2001) affect how evaluations are undertaken.

We are now at a transition point where the rhetoric or theory of development evaluation is moving from an accountability or audit approach to an effectiveness or learning approach. This transition is, however, neither seamless nor uniform. There is both a lag in the reorientation of evaluative systems and procedures, and a 'collision' of evaluative approach and models which highlight the contested purposes of development evaluation. Despite the rhetoric suggesting some preordained trajectory in development evaluation, the outcome of this contest over evaluative purpose in fact remains unclear to this point and future practice remains obscured.

Professionalization

Development evaluation is a nascent and evolving practice still in the process of professionalization. There have been major steps towards professionalization. Over recent years, OECD-DAC has exercised a leadership role in this professionalizing process. In 1991, it formulated the *DAC Principles for Evaluation of Development Assistance* which enshrine impartiality and independence, credibility, usefulness, participation and cooperation as the principles of development evaluation. They also embody five criteria which have been consistently applied in ensuing donor practice: relevance, effectiveness, efficiency, impact and sustainability (OECD-DAC, 1991). It then served a normative role by publishing a glossary of terms in evaluation and results-based monitoring (OECD-DAC, 2002). In 2006, it introduced *Evaluation Quality Standards* as a guide to good practice to harmonize the conduct of development evaluations, which were adopted in 2010. These *standards* relate to purpose and objectives, scope, methodology, information sources, independence, ethics, quality assurance, results, and completeness (OECD-DAC, 2010). These principles, criteria and standards have been broadly adopted by donors and now serve as the generally recognized benchmarks for development evaluation.

Other bodies contributing in the formalization of development evaluation at the global level include the United Nations Evaluation Group (UNEG), and the American Evaluation Association (AEA).

Professionalization is the process of formalizing a domain of occupational practice with the view to promoting and codifying the knowledge and skills expected of a profession. The current conflicts surrounding development evaluation is significant in explaining the continuing lack of established orthodoxy and the slow pace of professionalization. The recency of the evaluation field is illustrated by this journal which was launched in 1995 and in international development IDEAS (the International Development Evaluation Association) was only established in 2002.

From Paris to Accra – Improving development effectiveness

Over the past decade, the international community has repositioned evaluation at the centre of the development stage as a means of promoting development effectiveness and attaining the United Nations' *Millennium Development Goals* (MDGs). This followed a series of 'roundtables' at Monterrey (2002), Rome (2003), Marrakech (2004), Paris (2005), and most recently Accra (September 2008) and Doha (December 2008).

In 2002, the international community of 189 United Nations states, comprising both donors and developing countries, signed what has become known as the *Monterrey Consensus on Financing for Development*. This consensus addressed mounting concerns over dramatic shortfalls in resources required to achieve the internationally agreed development goals contained in the MDGs. It 'urged' developed countries to make concrete efforts towards the target of 0.7 percent of gross national product as overseas development aid to developing countries (United Nations, 2002). This consensus has been described as a 'watershed,' which created a new development paradigm framed on attaining the MDGs. This paradigm emphasized results, partnership, coordination and accountability, and introduced a new agenda which emphasized learning and continuous feedback at all phases of the development cycle (Picciotto, 2002b). Monterrey marked a new focus on managing-for-development-results (MfDR) which was consolidated at the *Rome Declaration on Harmonization* in 2003 and the *Second International Roundtable on 'Managing for Development Results'* in Marrakech in 2004 (OECD, 2003, 2004).

The pivotal roundtable was convened in 2005 to formalize the *Paris Declaration on Aid Effectiveness.* This Declaration embedded five principles as being core to ongoing development approach, hereafter known as 'the *Paris Principles*':

- *Ownership* – primacy vests in developing countries for leadership of policies and strategies and coordination of development actions;
- *Alignment* – donors base support on partner countries' national development strategies, institutions and procedures;
- *Harmonization* – donors' actions are harmonized, transparent and collective;
- *Managing for results* – all parties manage resources and decision making for improved results;
- *Mutual accountability* – all parties are accountable for results reaching targets set by the *Millennium Development Goals* (OECD, 2005).

A central feature of these principles is their emphasis on improving performance, i.e. ensuring that both government and donor activities achieve desired results. These results must now be demonstrable. This new emphasis reflects what is called 'managing-for-development-results' (MfDR).

Indicator 11 of this declaration, for example, requires results-oriented reporting and assessment frameworks that monitor progress against development strategies, and track a 'manageable number of indicators for which data are cost-effectively available'.

Millennium Development Goals

The imperative to attain the United Nations *Millennium Development Goals* has impelled the repositioning of development evaluation. In 2000, the MDGs were agreed by 189 countries as the overarching goal to reduce global poverty. The MDGs embody eight goals to eradicate extreme poverty and hunger, achieve universal primary education, promote gender equality and empower women, reduce child mortality, improve maternal health, combat HIV/AIDS and other diseases, ensure environmental sustainability, and develop a global partnership for development. These goals are supported by 18 measurable time-bound targets with 48 performance indicators which now provide the overarching focus of all development assistance. As such, they are a classic artefact of MfDR and are constantly monitored to gauge progress.

The significance of the MDGs at the pinnacle of international development doctrine cannot be overstated. They have undoubtedly galvanized global attention to meeting the challenge of poverty alleviation. While the MDGs have been useful at a universal level, their relevance in particular country contexts may be subject to debate. Additionally, the MDGs are not without their problems. Many hold serious misgivings about the MDGs as an effective development mechanism (United Nations, 2005). Some debate their conceptual integrity as an apex metric of poverty (Alkire, 2003). Others argue that their selectivity eclipses dimensions of inequality and is problematic in enabling politicians to turn a blind eye to the high, and generally rising, levels of inequality within their respective countries (Saith, 2006). It is also problematic that law and justice are not among these development goals, targets or indicators, other than by passing reference (United Nations, 2000). This has left governance and justice outside the big tent of measuring development, creating some of the challenges outlined in this article.

Despite these difficulties, the *Paris Principles* have unquestionably galvanized a global focus on delivering results to meet the MDG targets of 2015 and on developing the capacity to monitor their delivery within this timeframe. There is now a heightened emphasis on monitoring and evaluation to support delivering development results. This shift reflects the sentiment that aid had not sufficiently demonstrated its results in the past, and is now preoccupied with '*results, results, results and results*' (Van den Berg, 2005: 11).

Progress has, however, been both mixed and slow. In 2008, the OECD conducted a mid-point evaluation of implementation which found that unless countries 'seriously' gear-up efforts, the *Paris* targets to enable partner countries to achieve their own development goals would not be met (OECD, 2008). Its appraisal of implementation was sober: 'we are making progress, but not enough . . . [a]chieving development results – and openly accounting for them – must be at the heart of all we do' (United Nations, 2008).

The explanation for this slow progress is seen to be a need for evaluation capacity-building (MacKay, 2002). A growing number of governments and donors are working to create systems to promote and measure performance. But, there is as yet no 'best' model of what a government monitoring and evaluation system should be (MacKay, 2006). In practice, capacity-building is usually confined to training, when it should aim to improve the four dimensions of institutional, organizational, technological and human capacity-building (Schiavo-Campo, 2005).

The international community has refocused the rhetoric of development assistance on results, and on providing the means of monitoring and measuring those results. However, progress to date has been limited, even though this refocusing is potentially transformational.

Managing-for-development-results

As already noted, the notion of managing-for-development-results has been adopted by the international development community as its overarching means to assure improved development effectiveness.

MfDR is a management approach which aims at achieving important changes in the way that organizations operate by improving their performance. The OECD describes this approach as focusing the dialogue on results throughout the development process; by aligning programming, monitoring and evaluation with results; managing 'for not by results', and using results information for learning and decision making. MfDR provides management framework and tools to improve organizational learning and performance reporting (OECD and World Bank, 2008). OECD defines *results* as being the output, outcome or impact of a development intervention, and *results-based management* (RBM) as a management strategy focusing on performance and achievement of outputs, outcomes and impacts.

MfDR is hardly novel; it is an artefact of New Public Management (NPM) and in commercial settings has been labelled 'results-based management approach'. A key component of this approach is performance measurement, which is the process of objectively measuring how well an agency is meeting its stated goals or objectives. This approach rests on a construct which include identifying measurable objectives aided by logical frameworks, selecting indicators to measure progress towards each objective, setting explicit performance targets for each indicator, developing a performance monitoring system to regularly collect data on actual results, reporting on and reviewing those results, and integrating evaluations to analyse and advise on performance (Binnendijk, 2000). Other elements include baselines, indicators, data collection, stakeholder perceptions, systemic reporting, strategic partnerships and evidence of success/failure. Kusek and Rist (2004) propose a 10-step methodology to building, maintaining and sustaining a results-based monitoring and evaluation system.

Overall, there is consensus that MfDR places a distinctive emphasis on performance measurement which rests on three interrelated notions: the *logic model* of evaluation, called the 'logframe' at the operational level, which addresses the linkage between objectives and results; selection of *performance indicators* to reflect changes connected to an intervention, or to help assess the performance of a development actor; and review of *performance data* which requires the periodic collection and analysis of progress typically comparing actual results with planned targets.

Critiques of MfDR

A general critique of MfDR is that it does not necessarily accomplish what it sets out to achieve, i.e. provide an assured prescription for improving development effectiveness. This critique argues that MfDR is prone to weaken rather than strengthen the evaluation function. Its central tool, the 'logframe', is systematic but ungainly, torn by rival expectations to supply retrospective accountability and prospective learning. Additionally, requirements for independence and integration compete in operational practice with the effect that learning is largely excluded from the development management cycle. MfDR may be snake oil that cannot deliver what it promises!

The focus of MfDR on performance reporting is intended to strengthen accountability for delivering results. In this sense, it is consistent with the classic purpose of evaluation. But, there are a number of tensions which exist in how this accountability is provided.

First, the emphasis which MfDR places on the performance monitoring function implies a relatively straightforward complementarity with the associated evaluation function. But, at times, these functions compete. For example, when USAID (United States Agency for International Development) established its performance management and measurement systems in the 1990s, the size of USAID's central evaluation office staff and resources declined considerably, as did the number of evaluations conducted by its country operating units (Binnendijk, 2000). This illustrates that evaluation resources and activity may in practice decline − at least in some agencies − as a perverse consequence of the emerging pre-eminence of performance monitoring taking a larger slice from the M&E cake. Hence there is a discernible risk that MfDR may in fact divert finite funding and weaken the distinctive function of evaluation.

Second, MfDR places emphasis on the attainment of outcomes, results and impacts in what often becomes a bureaucratized practice of checking the 'logframe' and ticking the box at innumerable stages of project management. But human-centred change which is invariably complex and often unpredictable may not be amenable to this linear style of management. This is the case in assessing the qualitative dimensions of reforming justice where multi-dimensional political-economy factors are irreducible to the simplistic cause-effect relationships of a logframed approach. Demonstrating attribution is increasingly difficult for higher-order political economy outcomes, which then oblige managers to focus and report at lower results levels. This pits the accountability aspect of 'logframe' reporting, which is retrospective in orientation, against the improvement aspect of managerial planning, which is prospective. Each purpose requires different data and analysis, but the 'logframe' is invariably used simultaneously for both purposes and, by straddling both functions, becomes ungainly.

Third, in practice there is often a direct competition between the requirement for independence and impartiality which lie at the heart of the evaluation function, and the need to integrate an effective cycle of organizational learning at the operational level of both donors and developing governments. Balancing these competing interests is a growing concern and remains an unresolved challenge. Recent consolidation of the accountability function has been detrimental to developing effective learning systems, procedures and practices. Accountability reporting and management improvement compete to the point where, at the operational level, learning development lessons often remain systemically limited, usually in departmental silos. While it is acknowledged that independence does not invariably mean isolation, the evidence of practice highlights the challenge in balancing organizational systems to address both needs simultaneously.

The resurgence of performance measurement is an important feature of evaluation. However, earlier vogues of 'management by objectives' and 'management by results' which rested on similar notions in the 1970s and 1980s, disappeared because they did not serve their intended purposes. The cause of this failure lies in the audit-orientation of approach: counting only what can be counted results in the tendency for more attention to be given to issues of efficiency rather than effectiveness, diverting attention from what is really important, though more difficult to quantity:

> [M]easurement-based approaches do not of themselves necessarily lead to any increase in performance or a focus on outcomes . . . *but the tragedy is that they frequently appear to do so*. (Perrin, 1998: 377, italics added)

There is now mounting disquiet over MfDR. While results-based management places a high priority on measuring outputs and outcomes, evaluation studies demonstrate that there is no empirical correlation with results. Moreover, it generates unintended effects termed the 'performance paradox'. This refers to the weak correlation between performance and performance indicators caused by the tendency of indicators to become decoupled over time so that the relationship between actual and reported performance declines. Performance indicators can also lack neutrality, and become ambiguous and contestable. Additionally, performance indicators can unwittingly cause dysfunctional effects, driving managers to generate good reports as distinct from good outcomes (Van Thiel and Leeuw, 2002).

MfDR can also promote perverse consequences. For example, it encourages a systemic focus on easier more visible aspects of performance at the lower level of specific activities rather than at the higher level of consolidated policy. This creates the illusion of better performance with regard to controllable matters which are less societally significant. Similarly, measuring notions of quality, which is difficult and often contentious (Bouckaert and Peters, 2002). Others argue that the emphasis on monitoring builds on disillusionment with evaluation, the results of which are often not available at the time that decision making is required. Hence, there are distinct tensions between the monitoring and evaluation functions, which is described as a 'profound challenge' where investing in monitoring may actually replace investment in evaluation (Neilson and Ejler, 2008: 172).

In sum, the introduction of MfDR risks promoting monitoring at the expense of evaluation. We must remain on our guard against the illusion of developmental certainty which accompanies the new edifices of often meaningless, irrelevant or misleading performance data which are prone to being misunderstood and misused. There is now a heightened need for evaluation to reassert its relevance in the face of the encroachment of monitoring.

Arenas of debate

This critique extends to more particular analyses of the key elements of MfDR, notably its reliance on the logical framework and issues of methodology associated with evaluative models. It also opens the door to one of the most divisive debates in development evaluation: the 'paradigm war' between positivism and constructivism.

Project management and the logical framework approach

The logical framework is an all-pervasive tool of development planning and monitoring which, while clearly useful, risks oversimplifying the manifold challenges of development management. This is notably the case in qualitative and human-centred change arenas such as judicial reform where it is of quite limited utility for the purposes of monitoring and evaluation.

The logical framework approach is an integral element of MfDR. This framework, or 'logframe', was first introduced to development as a tool of project cycle management by USAID in the early 1970s. Prior to that time, project objectives were seldom clearly specified, and it was rarely clear what projects were specifically intended to achieve except in most general terms:

> Trying to evaluate (projects without objectives) was like trying to ride a bicycle with loose handlebars! (Cracknell, 2000: 94).

The 'logframe' is widely used by project designers to plan interventions (OECD-DAC, 2002), by implementers to guide their work, and by evaluators to assess actual achievements against the targeted indicators (Department for International Development, 2005). It comprises a matrix which deconstructs a project or programme into its component parts, namely *inputs,* resulting in activities, or *outputs*, together with the *risks* and *assumptions* involved and indicators of progress towards the achievement of *objectives*. The underlying logic of this approach is that if certain inputs are supplied and activities are undertaken, then intended outputs will result, given certain assumptions, which will make a contribution towards the achievement of certain wider objectives – sometimes termed vertical logic.

The conceptual rationale for this approach lies in what is often described by donors as a ‘results chain’ which links the intervention’s inputs to immediate outputs, and then to outcomes and final impacts or results. Mounting concerns over development effectiveness have emphasized the need to demonstrate the effects of an intervention, shifting the ultimate focus of both monitoring and evaluation from the development process to its results. Evaluations are then supposed to more readily focus on the efficacy of all or part of this results chain, and in practice are commonly concerned with appraising the pivotal relationship between ‘inputs’ and ‘outputs’.

The ‘logframe’ has become pervasively influential on the design, implementation and evaluation of the development activities of most, if not all, donors over the past 15 years under one name or another. Arguably, it is becoming more popular to donors across this transitional period, because it can be used irrespective of whether they are agitated by ‘audit’ or ‘effectiveness’ preoccupations. The ‘logframe’ is described by DfID (Department for International Development), Britain’s aid agency, as the most common planning, monitoring and evaluation tool used to chart the anticipated chain of cause and effect in development interventions. While this method has received criticism for its apparently rigid ‘blueprint’ approach to bringing about social changes, it has nevertheless predominated in the aid industry (Crawford et al., 2004).

The ‘logframe’ has many exponents who extol its use to clarify the objectives of a proposed intervention, imposing coherence in the expected causal links, and facilitating the monitoring through performance indicators at each stage. For these reasons, it is used heavily as both a planning and monitoring tool. Despite the ubiquity of this approach, there is, however, a quite startling lack of accompanying theory concerning its evolution or analysis of its application (Bell, 2000).

In my experience, the ‘logframe’ works much better in planning reform targets and monitoring their attainment than as an evaluative tool of managerial learning. However, it is a prosaically mechanistic approach to evaluation’s quest for the appraisal of value and worth and it has severe limitations in its ability to address any of the difficult questions which lie at the heart of the evaluative function. These questions usually relate to ascertaining and exploring the existence and nature of key relationships in complex change management processes because of the false assurance of linear change that it promises.

Notwithstanding their limitations, I have observed that ‘logframes’ have pervaded the planning and management of projects and have, in practice, assumed an overarching utility. While this tool usually only measures indicators of process and efficiency, it has allowed many actors at the operational level to focus on ticking the boxes on the ’deliverables’ in the ‘logframe’ as the day-to-day means of assuming developmental effectiveness. As a consequence, many donors rely on the logical framework in conjunction with rating scales for evaluation purposes. This has led to the often bizarre spectacle of donor evaluations cheerily asserting glowing project success, blinkered by the numbers in the boxes and oblivious or indifferent to the unchanged reality which surrounds the process on the ground (Savedoff and Levine, 2006).

At the heart of the problem, ‘logframes’ are seductively convenient over-simplifications of complex change management phenomena which distort development reality. Some observers have spoken of the *logicless frame* because an illusion of logic is created by the framework being superimposed at a technical level after the project has been formulated at the political level. Alternatively it is the *lackframe*, omitting critical dimensions in the change management process. Finally, it has been called the *lockframe* as the framework risks becoming fixed, out-of-date and irrelevant to any dynamic ongoing reform process (Gasper, 2000).

The fallacy of MfDR asserts that if Activity A is done, Output B will result. Among the most potent proponents of this critique, Ramalingam and Jones (2008) argue that development reality is characterized by its complexity and the dynamic of chaos. They criticise the prevalent development approach as being trapped in a paradigm of predictable, linear causality. The management of aid then becomes problematic because it denies or is oblivious to the reality of a world of messy and unpredictable change. This vision of change requires potentially radical political, professional, institutional and personal shifts for the development model to become effective. Logframes impose a reductionist model of understanding that assume a predictable and orderly universe of over-simplistic, linear cause-effect chains.

As a tool of management, logframes are too readily misused to imply that the sum of efficient inputs 'must', at some future time, lead to intended development results. My experience of practice indicates that it is never as simple – nor development change as linear – as the logic-model asserts.

Paradigm war over evaluative models

In addressing the question 'Does aid work?', the major debate centres on the evaluative model being used. This debate reveals the contested means of answering this question, and challenges the pre-eminence of the positivist approach to demonstrating impact and results. It also highlights the mounting dissatisfaction of many development professionals who are resorting to an altogether different model with which to assess merit and worth.

This central debate over evaluative modelling has been described by some commentators as a 'paradigm war'. This debate pits those who may be called *positivists* against *constructivists.* At its essence, it is concerned with the issues of how evaluation finds truth and contributes to knowledge. On the one hand, positivists advocate a traditional scientific approach. They are primarily concerned with establishing the validity and reliability of data, adopt experimental methods and counterfactual measurements and are preoccupied with the overarching need for methodological rigour. On the other hand, constructivists are primarily concerned to hear the voice of stakeholders, notably the alienated poor. They use participatory methods, case-studies and observations and refute the scientific approach as being costly, impractical and irrelevant. This debate has been described as a contest between the scientific, objectives-based project management model and the empowerment and 'pro-poor' stakeholder participatory model (Cracknell, 2000).

The outcome of this debate is very significant because it determines the overall approach to development evaluation, its methodologies and the nature of the data to be relied on.

Positivist approach to impact

Over recent years, the positivist approach has prevailed in development evaluation. This approach seeks to understand the social world by uncovering universal laws through the measurement of what David Hume called the 'constant conjunction of events' between two or more phenomena. These 'laws' are empirical generalizations which are seen to be mainly independent of time or space and are neutral and value-free. For positivists, the observation that two variables are strongly correlated is often understood to signify a causal relationship. Positivists discover empirical generalizations by setting up and testing hypotheses in a deductive manner, with hypotheses being extrapolated to a wider range of cases. In this respect, positivism is a form of naturalism, based on a belief in the unity of the natural and social sciences, and hence it attempts to replicate in social research the requirements and methods of the physical sciences, such as prediction, closed experimentation and the separation of research findings from interpretation (Prowse, 2007).

The attraction of positivism is its offer to provide a much-needed assurance of certainty to the business of development, which over recent years has become fraught with the lack of results. Burnside and Dollar are exemplars of the positivist approach; they used econometric methods to show that development aid does work in what they term a 'good policy environment', leading to faster growth, poverty reduction and gains in social indicators. Their groundbreaking empirical findings were based on historical cross-country regression analysis. They have been immensely influential in development policy, providing the empirical justification for the mantra on good governance which emphasizes that though aid can nurture development, the quality of spending is as important as its quantity (Burnside and Dollar, 2000; World Bank, 1998).

Positivism has been pervasively influential in MfDR through the realignment of the development discourse towards results and improving effectiveness. It promises to prove that aid really does work by scientifically demonstrating the counterfactual. This seeks to measure change over time by using either before/after measures or control groups to identify the effects of the intervention. Significantly, positivism raises the notion of impact evaluation to prominence by framing the counterfactual question: 'How would individuals who participated in the program have fared in the absence of the program?' (Duflo, 2003: 1), even though impacts are defined by OEDC as primary and secondary long-term effects, produced by a development intervention (OECD-DAC, 2002). At a more technical level, the World Bank defines impact evaluation as a counterfactual analysis of the impact of an intervention on final welfare outcomes where the counterfactual is a comparison between what actually happened and what would have happened in the absence of the intervention (White, 2006). The goal of impact evaluation is generally seen as attributing impacts to one project and to that project alone (World Bank, 2006).

Recognition of the evidentiary supremacy of the positivist position is, however, far from assured. Many argue that the empirical evidence is equivocal and that demonstrating a causal relationship between development policies and growth is fragile and often statistically insignificant. The ambivalence of the evidence is acknowledged by the World Bank at the technical level:

> Given that a number of the traditional arguments for why there might be a link between aid flows and economic growth appear to be *empirically weak* . . . it should come as some comfort that the majority opinion in the aid and growth literature appears to be that aid, *at least in certain circumstances*, can promote growth. (Kenny, 2006: 6, italics added)

From a positivist perspective, it may be argued that any evaluation which omits an assessment of impact and results might be seen as being deficient in terms of best practice in the post-*Paris* environment. Yet, the reality is that impact evaluations remain relatively rare, although there are a range of recent initiatives to expand their use (e.g. NONIE; J-PAL; 3iE). In essence, this rareness is because impact evaluations are relatively difficult, expensive and slow. While the counterfactual approach seems compelling, it has been argued that it is 'well nigh impossible' to set up a robust experimental control in development settings (Cracknell, 2000: 128). They may be theoretically rigorous, but are often not feasible in practice (Baker, 2000). Despite the billions of dollars spent on development assistance each year, there is still very little known about the actual impact of projects on the poor. Evaluations are often disproportionately expensive. The cost of impact evaluations is estimated at the World Bank in the order of US$300–500,000 which will exceed the total value of some niche interventions, though it is noted that IADB has reported conducting impact evaluation in Latin America at a lower cost. Other common difficulties, such as limitations in the quality and availability of reliable data and the shortage of available time to demonstrate results, may also explain why impact evaluations have never been undertaken in the area of law and justice (White, 2006).

These practical difficulties with applying positivist approaches is impelling practitioners to find alternatives which balance the search for evaluative knowledge with what is feasible in the development environment. A challenge for development evaluation is to realize its potential as a tool for learning − what works, what does not and the reasons why/why not − as well as institutional accountability. To bridge the gap between theory and practice, however, trade-offs are required between evaluation rigour and budget, time and data which are available in these 'real world' contexts to balance quality with feasibility (Bamberger et al., 2007).

Constructivism and participatory evaluation

Mounting dissatisfaction with the philosophic and conceptual deficiencies of positivism has impelled a search for different models of development evaluation as in evaluation more generally:

> In the past, the methodology employed in evaluations has been almost exclusively scientific, grounded ontologically in the positivist assumption that there exists an objective reality driven by immutable natural laws, and epistemologically in the counter assumption of a duality between observer and observed that makes it possible for the observer to stand outside the arena of the observed neither influencing it nor being influenced by it. (Guba and Lincoln, 1989: 12)

Constructivists assert that even when robust positivist evaluations are feasible, they are often either irrelevant or misleading:

> Over the past 30 years struggling with formal positivist evaluations, problems of comparability, unknown causal linkages and magnitudes, multiple causality, and the counterfactual, not to mention the costs and shortcoming of questionnaires and their analysis . . . (s)ome very experienced evaluators are now doubtful whether it is ever possible, with people-centred projects, to establish causal linkages between development interventions and possible outcomes. (Cracknell, 2000: 129)

Advocates of constructivism focus on negotiation, or what they term a 'hermeneutic dialectic', to determine what evaluation questions are to be asked and what information is to be collected. They also have a 'constructivist methodology' to carry out the inquiry process within the ontological and epistemological presuppositions of the constructivist paradigm (Guba and Lincoln, 1989: 83−4).

Beneath the verbiage, this model is important because it reframes the discourse. It refutes what it sees as the false assurance of positivist evaluation and provides a coherent alternative. They advocate a paradigm of evaluation criteria which substitutes *credibility* for internal validity, *transferability* for external validity, *dependability* for reliability, and *confirmability* for objectivity (Guba and Lincoln, 1989: 249). This model acknowledges an asymmetry of power in the evaluation transaction which is seen as being 'the great unmentionable in evaluation theory' (Cracknell, 2000: 88). It redresses this asymmetry through a shift to participatory monitoring and evaluation by local people.

Participatory evaluation is a paradigm shift. It is as an evaluation approach in which representatives of agencies and stakeholders, including beneficiaries, work together in designing, carrying out and interpreting an evaluation (OECD-DAC, 2002). Significantly, this paradigm follows from Guba and Lincoln's challenge that by staying detached, the evaluator merely ensures that s/he can never really understand what is going on in terms of the beneficiaries themselves and their aspirations. In effect, the evaluator 'must' become involved and participate. While this model directly conflicts with positivistic notions of objectivity, detachment and impartiality, it has received much endorsement. Some support this shift as 'a giant leap into new territory' on the basis that classic notions of neutrality are a fiction (Cracknell, 2000: 333). Patton goes further: '[t]o claim the mantle of objectivity in the post-modern age is to expose oneself as embarrassingly naïve' (2002: 50).

Many in the development community are moving towards notions of participatory evaluation and stakeholder analysis which recognize that implementing a people-centred development project is an iterative continuous process more concerned with learning than accountability. This recognition conceives the change process as being flexible and needing to be adaptive, rather than linear and rigidly directed at targets. This shift involves notions of empowerment, social democracy and popular participation, requiring fundamental changes in approach including 'shifts of role from evaluator to facilitator, of style from judgment to learning, of mode from extractive to empowering, and of focus from one-off to ongoing process' (Cracknell, 2000: 27).

While donor rhetoric increasingly reflects participatory notions, the actual practice of participatory evaluation, however, remains limited. This may be because there is still little uniformity of understanding about what participatory evaluation involves or how it will be put into practice:

> The disturbing thing is that for all their brave words in favour of more participatory methods of evaluation, most donor agencies are still conducting their evaluations in a way that very much diminishes the possibility of this actually happening. (Cracknell, 2000: 320)

The significance of this core debate for the current discourse of development evaluation is that the positivistic approach – with its emphasis on science and linearity – may be seen as fitting well within the MfDR/NPM/RBM models, outlined above. The methodological challenges and costs associated with impact evaluation are probably a reason why more simplistic monitoring models are crowding out evaluation.

Conclusions

I have argued that development evaluation is a young, still evolving – hence adolescent – discipline, which is characterized by a number of major debates and unanswered questions. Over the past decade, OECD-DAC and professional evaluators working in national and international associations such as IDEAS has provided concerted leadership to professionalize evaluation concepts, principles, norms and standards. Significant challenges, however, remain before this endeavour will attain the capacity and standing to meet the needs of the international community to assess development effectiveness.

For the present, development evaluation is internally contested over fundamental aspects of its purpose and approach; substantial technical challenges exist in demonstrating causality and attribution; divisive, unresolved disciplinary debates simmer over methodology and rigour; and practice is consequently fragmented by competing orthodoxies over independence and integration.

In the meantime, development has entered a new era of results-based management. This has repositioned the tools of monitoring and evaluation to the centre of the development stage. The shift in focus to measuring effectiveness embodied in the *Paris Principles* involves a profound transition for both donors and developing countries.

I draw four main conclusions from the above discussion:

- First, development evaluation has become significantly more important as a key tool in implementing the international consensus to improve development results. Mounting disillusionment at the perceived failure of aid to redress the 'poverty gap' has galvanized a major transition in development evaluation, resulting in substantial growth and refinement, and heightened emphasis on demonstrating effectiveness through monitoring and evaluating for impact.
- Second, in practice this 'transition' has faced, a range of substantial challenges, many of which are yet to be resolved. Development evaluation remains in an unresolved state which has slowed down its emergence as a bounded professional discipline. This is evident in the paradigm war over the choice of evaluative models which has riven the discourse between positivists and constructivists.
- Third, there is an unresolved tension between the accountability and learning dimensions of development evaluation which may be irreconcilable. This may explain the current trend towards performance-based models which markedly emphasize monitoring at the expense of evaluation.
- Fourth, *managing-for-development-results* has emerged as the new development doctrine in the post-*Paris* setting which embodies a shift in focus from monitoring the efficient delivery of outputs to monitoring impact and evaluating the effectiveness of results. To date, impact evaluations remain rare owing to technical difficulty in development environments and their disproportionate cost. The seductive convenience of the 'logic model' to steer all aspects of change risks being reductionist and over-simplifying the evaluation of nuanced, complex and non-linear human-centred change characteristic.

Addressing these challenges – only some of which are transitional in their nature – is required now, but will require time. Some, such as the realignment of donor systems for monitoring and evaluating effectiveness of development endeavour are relatively straightforward and may be expected shortly, though domestic political demands and institutional fiscal constraints render it unlikely that accountability-based systems will ever be dispensed with by donors altogether. Others, such as implementing the new paraphernalia of MfDR, will clearly require more time to design frameworks and collect data which are sufficient for trend analysis and the like. Yet others, such as the simmering disjunctions of the 'paradigm war', shall persist behind uneasy interdisciplinary truces. In the meantime, the expectations of the international community for evaluation to deliver practical insights in promoting development effectiveness, notably relating to the MDG's, are likely to intensify in the near term rendering research on and consideration of these unresolved issues all the more topical.

While there has been an undoubted shift or transition in development evaluation over the past decade, it is incomplete and unresolved to this stage. Until these challenges are addressed, there will remain a marked '*evaluation gap*' between the theory and rhetoric of the *Paris Principles on Aid Effectiveness* and the real world of development evaluation practice.

References

Alkire S (2003) A conceptual framework for human security. Centre for Research on Inequality, Human Security and Ethnicity, Oxford, Working Paper, No. 2.

Armytage L (1996) *Educating Judges*. London/Boston: Brill/Kluwer Law.

Baker J (2000) *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: World Bank. URL (consulted 25 November 2009): http://go.worldbank.org/8E2ZTGBOI0

Bamberger M, Rugh J and Mabry L (2007) RealWorld Evaluation: Working Under Budget, Time, Data and Political Constraints. Professional Development Workshop 10, 6 November. URL (consulted 16 February 2011): http://www.realworldevaluation.org/RealWorld_Evaluation_resour.html

Bell S (2000) Logical frameworks, Aristotle and soft systems: a note on the original values and uses of logical frameworks in reply to Gasper. *Public Administration and Development* 20: 29–31.

Binnendijk A (2000) *Results-based Management in Development Cooperation Agencies: A Review of Experience*. Paris: OECD-DAC.

Bouckaert G and Peters G (2002) Performance measurement and management: the Achilles' heel in administrative modernization. *Public Performance & Management Review* 25(4): 359–62.

Burnside C and Dollar D (2000) Aid, policies, and growth. *The American Economic Review* 90(4): 847–68.

Cracknell B (2000) *Evaluating Development Aid: Issues, Problems and Solutions*. London: SAGE.

Crawford P, Perryman J and Petocz P (2004) Synthetic indices: a method for evaluating aid project effectiveness. *Evaluation* 10(2): 175–92.

Department for International Development (2005) *Guidance on Evaluation and Review for DFID Staff*. URL (consulted 16 February 2011): http://www.dfid.gov.uk/Site-search/?q=guidance+evaluation+2005

Duflo E (2003) *Does Development Aid Make a Difference?* Washington, DC: Center for Strategic and International Studies.

Gasper D (2000) Evaluating the logical framework approach: towards learning orientated development evaluation. *Public Administration and Development* 20: 17–28.

Guba E and Lincoln Y (1989) *Fourth Generation Evaluation*. Newbury Park, CA: SAGE.

J-PAL (n.d.) URL (consulted 16 February 2011): http://www.3ieimpact.org/userfiles/file/3ie_Leaflet_for_web.pdf

Kenny C (2006) What is effective aid; how would donors allocate it? World Bank Policy Research Working Paper, No. 4005.

Kusek J and Rist R (2004) *Ten Steps to a Results-based Monitoring and Evaluation System*. Washington, DC: World Bank.

Leeuw F and Furubo J (2008) Evaluation systems: what are they and why study them? *Evaluation* 14(2): 157–69.

MacKay K (2002) The World Banks ECB experience. *New Directions for Evaluation* 93: 81–99.

Mackay K (2006) Institutionalization of monitoring and evaluation systems to improve public sector management. World Bank Independent Evaluation Group and Thematic Group for Poverty Analysis, Washington, DC, Working Paper, No. 15.

Nielsen S and Ejler N (2008) Improving performance? Exploring the complementarities between evaluation and performance management. *Evaluation* 14(2): 171–92.

Norton J (2001) International financial institutions and the movement towards greater accountability and transparency: the case of legal reform programs and the problem of evaluation. *International Lawyer* 35: 1443–79.

NONIE (n.d.) URL (consulted 16 February 2011): http://www.worldbank.org/ieg/nonie

OECD (2003) *Rome Declaration on Harmonisation* URL (consulted 16 February 2011): http://www.oecd.org/dataoecd/54/50/31451637.pdf

OECD (2004) Second international roundtable on 'Managing for Development Results' 2004. Marakech: DCD/DIR.

OECD (2005) *Paris Declaration on Aid Effectiveness: Ownership, Harmonisation, Alignment, Results and Mutual Accountability*. URL (consulted 16 February 2011): http://www.oecd.org/dataoecd/11/41/34428351.pdf

OECD (2008) *Survey on Monitoring the Paris Declaration – Effective aid by 2010? What will it take?* Paris: OECD.

OECD (n.d.) *DAC Criteria for Evaluating Development Assistance*. URL (consulted 16 February 2011): http://www.oecd.org/document/22/0,2340,en_2649_34435_2086550_1_1_1_1,00.html

OECD-DAC (1991) *Principles for Evaluation of Development Assistance*. Paris: OECD.

OECD-DAC (2002) *Glossary of Key Terms in Evaluation and Results-based Monitoring*. Paris: OECD.

OECD-DAC (2010) *Quality Standards for Development Evaluation*. Paris: OECD. URL (consulted 28 February 2011): http://www.oecd.org/dataoecd/55/0/44798177.pdf

OECD and World Bank (2008) *Emerging Good Practice in Managing for Development Results: Source Book*. Washington, DC: World Bank.

Patton M (2002) *Qualitative Research and Evaluation Methods*, 3rd edn. Thousand Oaks, CA: SAGE.

Perrin B (1998) Effective use and misuse of performance measures. *American Journal of Evaluation* 19(3): 367–79.

Picciotto R (2002a) The logic of renewal: evaluation at the World Bank (1992–2002). Speech delivered at the OED: The First 30 Years, Washington, DC, 23 September.

Picciotto R (2002b) Development cooperation and performance evaluation: the Monterrey challenge. Washington, DC: World Bank Operations Evaluations Department Working Paper.

Prowse M (2007) Aid effectiveness: the role of qualitative research in impact evaluation. Overseas Development Institute, London, Background Note, December 2007.

Ramalingam B and Jones H (2008) Exploring the Science of Complexity. London, Working Paper, No. 285.

Saith A (2006) From universal values to millennium development goals: lost in translation. *Development and Change* 37(6): 1167–99.

Savedoff W and Levine R (2006) When will we ever learn: improving lives through impact evaluation. Report of the Evaluation Gap Working Group, Centre for Global Development, Washington, DC.

Schiavo-Campo S (2005) ECD: building country capacity for monitoring and evaluation in the public sector: selected lessons of international experience. World Bank, Operations Evaluation Department, Washington, DC, Working Paper, No. 13.

United Nations (2000) *United Nations Millennium Declaration*, GA Res 55/2, UN GAOR, 55th Session, Suppl. no. 49, UN Doc. A/55/49. URL (consulted 16 February 2011): http://www.un.org/millennium/declaration/ares552e.htm

United Nations (2002) Monterrey Consensus of the International Conference on Financing for Development. In: *Report of the International Conference on Financing for Development*, Res 1, UN Doc A/CONF.198/11.

United Nations (2005) *Human Development Report 2005*. New York: UNDP, 40.

United Nations (2008) Third high level forum on aid effectiveness. Accra Agenda for Action, Ghana, Ministers' Statement, 4 September 2008.

United Nations Evaluation Group (2005) *Norms for Evaluation in the UN System*, 6. URL (consulted 16 February 2011): http://www.uneval.org/normsandstandards/index.jsp?doc_cat_source_id=4

Van Den Berg R (2005) Results evaluation and impact assessment in development co-operation. *Evaluation* 11(1): 27–36.

Van Thiel S and Leeuw F (2002) The performance paradox in the public sector. *Public Performance & Management Review* 25(3): 267–81.

White H (2006) Impact evaluation: the experience of the independent evaluation group of the World Bank. World Bank, Washington, DC, MPRA Paper, No. 1111.

World Bank (1998) *Assessing Aid: What Works, What Doesn't, and Why*. New York: Oxford University Press.

World Bank (2006) *Impact Evaluation and the Project Cycle*. Washington, DC: World Bank, Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation.

3iE (n.d.) URL (consulted 16 February 2011): http://www.povertyactionlab.org/about-j-pal

Dr. Livingston Armytage is Director of the Centre for Judicial Studies, Sydney (www.educatingjudges.com). He is a development specialist in justice reform and evaluation. Monographs include *Educating Judges* (Kluwer/Brill, 1996); *Searching for Success in Judicial Reform* (Oxford University Press, 2008), and *Reforming Justice: A Journey to Fairness* (Cambridge University Press, in press 2011). [email: larmytage@yahoo.com]