Experimentalism and development evaluation: Will the bubble burst?

Robert Picciotto

Evaluation 2012 18: 213
DOI: 10.1177/1356389012440915

Opinion piece

Evaluation
18(2) 213–229

DOI: 10.1177/1356389012440915
evi.sagepub.com
SAGE

Experimentalism and development evaluation: Will the bubble burst?

Robert Picciotto
King's College, London

Abstract
Bridging the current divide of opinion about experimentalism would help protect an evaluation brand currently under threat in international evaluation circles. In order to help settle a lingering and unnecessary controversy, this opinion article describes the policy force field that triggered the recent surge of interest in experimental methods in development evaluation; digs up the historical and philosophical roots of the long-standing epistemological debate; outlines the value and boundaries of experimentalism; and speculates about its prospects in development evaluation.

Keywords
development, economics, evaluation, experiments, history, quality, randomization

Introduction

The sudden popularity of experimental methods in development evaluation is counterintuitive. Whereas these methods are best suited to the assessment of simple and stable programs, the development enterprise is mostly made up of complex, adaptable interventions implemented in volatile environments. Why then did the 'paradigm wars' that were laid to rest within the mainstream evaluation community decades ago suddenly resurface in the development cooperation domain? What explains the virulence of the debate? What stands in the way of a rapprochement between the 'randomistas' and their detractors? What is the future role of experimental methods in development evaluation?

The assault on traditional development evaluation

The World Bank and other aid agencies have long produced reports that rate the aggregate success of their development interventions. For decades these documents escaped academic scrutiny and

Corresponding author:
Robert Picciotto, King's College London, Strand, WC2R2LS, UK.
Email: r.picciotto@btinternet.com

put forward success rates ranging from two-thirds to four-fifths of funded projects. The validity of these claims is now challenged by research economists who argue that the traditional rating methods used by development institutions to ascertain success lack rigour since they do not use a counterfactual to attribute results to the financed intervention.

Disappointing policy-research results

The onslaught on traditional development evaluation occurred at the turn of the century – a consequence of the internal strife that erupted within the development economics establishment. A cottage industry of policy-research studies grounded in cross-country regressions had generated diverse and contradictory findings regarding the impact of aid (Tarp, 2009). They could not identify robust correlations between aid volumes and economic growth. The often contradictory findings reflected the methodological limitations of cross-country correlations. The research designs:

- neglected the technological and capacity building benefits of aid;
- failed to distinguish between aid channels, instruments or modalities; and
- did not take account of the diversity of social and institutional contexts.[1]

The ambiguity of the results contributed to a growing public despondency about aid-effectiveness research. They also induced economists to raise pointed questions about the validity of claims advanced by in-house evaluation units. A literature seeking to explain the discrepancy between countrywide aid impacts and project-level studies (labelled a 'micro–macro paradox') emerged. It stressed the unintended, covert and perverse effects of aid on economic development.

The indeterminacy of policy-research results also contributed to a polarized intellectual environment that allowed two warring factions – the aid optimists led by Columbia University professor Jeffrey Sachs (e.g. 2005) and the aid pessimists led by William Easterly of New York University[2] – to engage in extended ideological battles that produced more heat than light and undermined public trust in development assistance.

The MIT upstarts

The disillusion associated with macroeconomic aid-effectiveness research opened up a strategic opportunity for young economists based at the Massachusetts Institute of Technology (MIT). They shifted the focus of the aid effectiveness debate from the abstract plane of macroeconomics to the gritty playing field of microeconomics. Staying clear of grandiose generalizations, they championed a fresh approach focused on a clinical examination of specific development interventions. Their research sought to figure out whether current development intervention models worked 'on the ground'.

This targeted approach to development economics postulated that the effectiveness of development assistance was best ascertained through experimental methods. The battle cry of the new development economists was sounded by the MIT Poverty Action Lab's charismatic co-founder (Esther Duflo) when she famously declared during a World Bank Conference on the evaluation of development effectiveness held in 2003: 'Just as randomized evaluations revolutionized medicine in the 20th century, they have the potential to revolutionize social policy during the 21st' (Duflo and Kremer, 2005).

A call to arms

While this bold proposition was challenged by eminent research economists (Deaton, 2005; Ravallion, 2005), it garnered the enthusiastic support of international philanthropic foundations intent on making their mark on the development scene. With financial help from the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation, an Evaluation Gap Working Group was assembled by the Center for Global Development (CGD) in 2004. Its underlying rationale was that billions of dollars and thousands of aid programmes had been devoted to improve health, education and other social sector outcomes without studies that could determine whether or not they actually 'worked'.

The Working Group Report (Center for Global Development, 2006) brushed aside the rating system used by development evaluators to assess the effectiveness of aid interventions. It asserted that the results of traditional evaluations lacked validity since they did not address the attribution question in a rigorous fashion. A systematic search for soundly based evidence about the effectiveness of development interventions was needed. It would help close ineffective programmes and identify approaches to poverty reduction worthy of replication. The report asserted that ascertaining whether aid worked required randomized field experiments or quasi-experimental methods that approximate the randomization gold standard.[3]

Never mind that this far reaching proposition had been dissected and found wanting decades earlier by large segments of the evaluation community. Evidently the lessons of past evaluation debates had not been internalized by the economics profession so that the momentum generated by the MIT economists proved unstoppable. Gradually research funding shifted from macroeconomic studies to microeconomic assessments of development interventions.

A twilight struggle

It did not take long for the struggle for dominance within the aid-effectiveness research establishment to spill over onto the development-evaluation scene. Aid evaluators who had only recently joined the mainstream of the evaluation profession were caught unawares. They had not been a party to the evaluation 'paradigm wars' of the late 1970s and early 1980s.[4] Unprepared for the onslaught they had to give ground.

By 2011 the MIT Poverty Action Lab had built an impressive coalition of 59 professors selected from leading universities. It had trained 851 people and it managed 144 evaluations in 43 countries. Similar units were created in other elite universities. This then is how economists and econometricians invaded a territory that had previously been the preserve of development evaluators wedded to cost–benefit analysis and/or participatory approaches.

It did not take long for noisy controversies to erupt in international evaluation conferences and for a schism to threaten the development evaluation community. At one end of the spectrum, seasoned development evaluators schooled in qualitative methods viewed the rigour attributed to experimental methods as largely illusory. At the other end, evaluators who had long sought closer connections with the social science disciplines welcomed the new economists' forays into the evaluation field and advocated close collaboration with them.

In a concerted effort to settle the conflict, a global Network of Networks for Impact Evaluation[5] was assembled. It brought together the evaluation units of all official aid agencies as well as representatives of evaluation associations under the aegis of the World Bank and the Development Assistance Committee of the Organization for Economic Cooperation and Development.

An uneasy truce

Following extended deliberations, broad agreement was reached on a methodological guidance document (Leeuw and Vaessen, 2009). It acknowledged the superiority of experimental designs to ascertain attribution in some circumstances but it did not endorse the hypothesis according to which randomized control trials (RCTs) constitute a gold standard. Instead, it favoured mixed methods adapted to the unique needs of specific evaluations.

This finely balanced posture, now widely endorsed, inaugurated an uneasy truce among the contending parties. But in truth, just as in the partial settlement that ended the paradigm wars within US evaluation circles, the methodological conflict had been frozen rather than resolved. Even now there remain fundamental differences of views within the development-evaluation community regarding experimentalism.

This divide of opinion matters. Attribution is a core mandate of the evaluation discipline and trust in the development-evaluation brand is shaken when eminent evaluation leaders express sharply different views about the fulfilment of a critical evaluation function. The task of consolidating a broad-based understanding about the precise role of experimentalism in development still lies ahead.

Fundamentalism versus pluralism

Randomization fundamentalists argue that experimental designs are the *only* scientific basis of ascertaining causation or attribution. Such an extreme position is untenable since biology, geology, astronomy, epidemiology, the forensic sciences, etc., all testify to the proposition that causation can be established without RCTs. Similarly, investigatory techniques, contestability protocols and rules of evidence (rather than controlled experiments) are widely considered sufficiently rigorous to penalize, jail and in some jurisdictions execute individuals convicted of a serious crime.

Randomized designs are not flexible enough to embrace the complexity of development contexts and interventions. Open-ended questions and qualitative approaches are better suited to address the diverse evaluation questions raised by evaluation commissioners and stakeholders. But fundamentalists are true believers. They enjoy moral certitude and do not accept evidence that contradicts the revealed truth. They are not persuaded by logical arguments. They exclude other perspectives, only associate with other believers and overcome resistance by non-believers through exclusion and compulsion rather than persuasion and open debate.

A defining characteristic of fundamentalism is that the only source of legitimate truth lies in the past. Fundamentalists frequently refer to sacred texts and sacred figures. The paradigm wars that have plagued evaluation are frequently waged by drawing authority from the historical past and it is therefore relevant to examine dispassionately the antecedents of the current methodological controversy. By adding perspective to the debate, the historical overview that follows may help promote mutual understanding and make a convergence of views more likely.

Origins of the debate

Evaluation as a practice has been around from time immemorial. Emperors, kings and princes have always sought to use the brain power of thinkers and scholars to assess the solidity of their rule, the loyalty of their subjects or the competence of their subordinates. Thus, the civil service of ancient Egypt evaluated the harvests of the Nile delta; entrance examinations were a regular feature of government in ancient China, etc.

Critical inquiry came into the picture with the ancient Greeks. Thales' rejection of mythological explanations and his articulation of explicit hypotheses were core ideas of the scientific revolution. Equally the history of science in China is long and rich and the Islamic civilization helped to lay the foundations of experimental science. But the institutionalization of scientific inquiry only came into its own in Europe at the beginning of the modern era. This is when experimentalism became a basic tenet of the scientific method.

Spiritual roots

The experimental tradition has deep spiritual roots. This may explain why debates about evaluation methods often evince passion. Experimentalism would not have become ascendant had natural investigation not been celebrated as a revival of innocent religion. Thus, through systematic reconsideration of biblical texts John Milton and his contemporaries provided new interpretations of the Creation. Their reformist conception of religious faith gave respectability to experimentalism (Picciotto, 2010).

Specifically, appeal to divine sanction validated a basic tenet of the scientific method according to which positive verification is the only authentic test for knowledge creation and accumulation. This principle led to a fundamental reconfiguration of the relationship between religion, experimental science and the public sphere. Thus, for Francis Bacon and his Royal Society disciples, observation uncorrupted by dogma was legitimized by the advent of a new strain of Christian apologetics that instructed the general public as well as scientists and scholars to secure evidence of divine wisdom through direct scrutiny of the natural order.

Eventually, positivism extended the same approach to human society by asserting that for the social sciences as for the physical sciences only knowledge that is testable, cumulative, transcultural and independent of the observer is valid. This doctrine was adopted without reference to any deity. But even then the sacred features of experimentalism resurfaced: in his later years Auguste Comte, the founder of sociology, developed a 'religion of humanity' inspired by positivist principles.

Philosophical roots

The torch that had been lit by Auguste Comte was passed to John Stuart Mill. The notion that scientific rationalism can illuminate human conduct is rooted in the British empiricism of John Locke, George Berkeley and David Hume as well as the utilitarianism of Jeremy Bentham. But it might not have received broad-based intellectual support without J.S. Mill's logical descriptions of inductive methods and compelling expositions of experimental inference rules.[6]

His sixth and final book, *System of Logic*, offered an influential account of social-science methodology buttressed by his own practice as a leading political economist and as a pioneer in the emerging discipline of sociology.[7] Next, Durkheim endorsed the notion of a natural science of human beings. But unlike J.S. Mill, he believed that sociology would have to create its own distinctive approach rather than replicate the methods of the natural sciences.

Max Weber further distanced himself from the positivist tenet according to which invariant generalizations about human relationships can be asserted outside a specific cultural context. He argued that given the complexity of human interactions, the social sciences could only uncover causal relationships among hypothetical simplifications (i.e. ideal models) of social phenomena.

The gap between the social and the natural sciences was widened by critical theorists and historical materialists such as Karl Marx, Theodor Adorno and Jürgen Habermas. Their competing theories converged on the proposition that the natural and social sciences are ontologically distinct.

By contrast, Thomas Kuhn discovered affinities between the two knowledge domains when he expressed the view that scientific theory choice depends on paradigmatic considerations that go beyond observation and logic.

The postmodern critics that followed went further. They debunked the scientific method altogether on the grounds that social facts are mediated through human consciousness so that all experimentation is subjective and even retrograde when it concerns society since value-laden, interest-driven interpretations of human reality inevitably intervene (Rosenau, 1991). Probing the interface between power and knowledge, these pioneers of the postmodern movement promoted social inquiry geared to communicative action in the public sphere, emancipation and social change.

Inevitably this advocacy orientation left the postmodernists open to sharp criticism and charges of subjectivity and bias. But, by that time, deep scepticism about evaluative claims that do not make their social purpose explicit had become widespread and positivism, especially in its utopian form, had lost its lustre. Science was no longer perceived as the ultimate arbiter of social policy and belief in human progress inevitably fuelled by technological development no longer held sway.

This said, 'value-free' experimentalism, while besieged from all sides, is still remarkably influential. In particular, quantitative, 'scientifically-based' research retains considerable prestige among policy makers and within the economics establishment. Deep cleavages within the social sciences and in evaluation are likely to persist as long as the limits of the scientific method within the social realm remain blurred.

Evaluation roots

The systematic use of observation to understand and guide social change is at the core of the evaluation discipline. The deliberate extension of the scientific method to the realm of public policy may be traced all the way back to 1662 when William Petty, a physician, joined the Royal Society and wrote a treatise on taxation. This pamphlet inaugurated a new art of 'political arithmetic' that Petty subsequently applied to the improvement of living conditions in Ireland.

Petty's pioneering contributions to economics and statistics presaged the Enlightenment era, when the quest for knowledge was further extended to legal regimes and political institutions. The seminal writings of Condorcet (produced between 1774 and 1794) showed the way. Condorcet rejected divine revelation and embraced empirical and rational inquiry as did all Enlightenment philosophers.[8] Emblematic of the Age of Reason, Condorcet's writings probed the art of government, the mathematics of democracy and the management of social change.

His ideas embodied conceptions of human progress that still resonate today. They were imbued with the belief that universal human values are compatible with freedom of thought and individual liberty. These ideas spread throughout Europe so that the application of independent reasoning to the assessment of government actions came into use (e.g. when government appointed commissions were set up in Sweden in the 17th century).

These historical antecedents notwithstanding, evaluation as a free-standing academic discipline focused on government programmes only flourished in the mid-20th century. It arose out of the ashes of the Second World War – a period characterized by 'can do' attitudes and trust in government. This is the rich intellectual soil ploughed by Marvin C. Alkin in his search for the roots of evaluation theory.

According to Alkin, all evaluation doctrines currently on offer can be classified by the extent to which they focus on methods, uses or valuing (Alkin, 2004). He metaphorically displays them as the three main branches of a bushy evaluation theory tree. Experimentalism occupies a privileged

position at the very base of the methodological branch: the notion that rigorous assessments of social programmes can make a major difference in people's lives was present at the creation of the evaluation discipline.

Specifically, evaluation pioneers concerned with social programmes conceived evaluation as a transmission belt between the social sciences and decision makers.[9] Donald Campbell, the visionary methodologist of the *Experimenting Society*, visualized public interventions as policy experiments. Sharply focused on the elimination of bias in social-science inquiry he promoted randomized tests as the methodological gold standard. Indeed, he even went so far as to tout the experiment as 'the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition' (Campbell and Stanley, 1963).

From thesis to antithesis

Campbell eventually came to his senses and reconsidered his views. Faced by the disappointing results of experimentalist studies he revised his negative assessment of qualitative methods and recognized that the identification and elimination of potential claims to causality and the interpretation of side effects of public interventions inevitably require expert qualitative judgement so that in order 'to be truly scientific, we must re-establish the qualitative grounding of the quantitative' (Campbell, 1974).

Thomas Cook built on Campbell's ideas by focusing on contextual factors and how they affect classical experiments. He developed quasi-experimental techniques designed to overcome difficulties associated with experimental control. He also stressed the importance of consultation with evaluation stakeholders. Similarly, Peter Rossi and Carol Weiss while recognizing the attractiveness of controlled experiments to eliminate selection bias, made seminal contributions to the methodological field by linking the programme logic underlying public interventions to theory-driven evaluations.

Lee J. Cronbach's intellectual journey led him even further away from a wholesale commitment to randomized field tests. It culminated in a fulsome rejection of classical experimentalism. Ultimately, Cronbach came to the view that only simplistic 'go/no go' decisions are influenced by randomized tests whereas the provision of useful evaluation data for instrumental use requires the exploration of a broad range of relevant issues rather than a narrow focus on the necessarily restricted set of questions amenable to RCTs.

Eventually Cronbach's interest in enlightened policy making through evaluation led him to question the external validity of RCTs. He ended up doubting whether robust generalizations about human behaviour can be secured through social research and he advocated modesty and restraint in the formulation of policy recommendations (Cronbach, 1982). Similarly, Robert Stake, who started as a positivist and a mathematician, became increasingly disenchanted with the potential of measurement and formal modelling in the assessment of social programmes.

Not yet an end to evaluation history

The trajectory of evaluation ideas sketched above confirms that once tested in the real world of policy practice, experimental doctrines inevitably yield ground to qualitatively-oriented approaches reliant on pluralistic, interactive and flexible methods. Thus, it is tempting to interpret the pioneers' evolving conceptions of evaluation as a dialectical movement in which experimentalism represents an original thesis that, once it proves inadequate, is superseded by its antithesis. But such a Hegelian

narrative is complete only when the flaws of the antithesis also come to light, thus opening space for a synthesis that reconciles the opposing doctrines.

Unfortunately, no cogent evaluation school can yet lay claim to the privileged status of an ideal synthesis. The mixed-methods doctrine shows great promise but it is still in its infancy and it has yet to develop a coherent narrative and a distinct set of evaluation practices. But one firm conclusion emerges: the restoration of experimentalism in development evaluation is not an end-state. The surge of interest in experimental methods within the international domain represents the temporary reversal of a trend rather than the ultimate result of a natural intellectual progression. The evaluation theory tree is still growing.

What then explains the remarkable ascent of experimentalism in evaluations of international-development programmes? Policy ideas tend to reflect the dominant concerns of society and it is no accident that evaluation methods that promise certainty should have gained influence at a time when the development community faces unprecedented challenges. Having probed the origins of the current experimental debate within development-evaluation practice, it is time to examine its proximate causes. They are part and parcel of the current development-assistance predicament. But they are also to be found in weak quality-assurance arrangements within the development-evaluation profession.

Weak quality assurance

A pervasive quality problem plagues evaluation practice in international development. Evaluations carried out by in-house evaluation units rarely address problems of selection bias. Poor 'evaluability' of development interventions goes a long way in explaining why this is so. Several meta-evaluations have revealed the astonishing neglect of rigorous attribution analysis in international development practice (Jerve and Villanger, 2008).

Nor is there any doubt that there has been underinvestment in RCTs even in cases where they are fully appropriate (White, 2009). The challenge of attribution posed by the economics profession highlights the fact that the development-evaluation community has yet to adopt, let alone enforce, uniform good-practice standards across borders. Credible methods geared to ascertaining causality in a rigorous way are now at a premium – and rightly so. On the other hand, Esther Duflo's claim that current medical-research practice represents a standard of excellence for the design of sound social policies is highly questionable.

The lure of medical research

Embarking on a social-transformation initiative through development aid is not the same as administering a pill. This is not to say that scientific work cannot achieve rigour in medical research or that randomization is not the method of choice to assess attribution in some circumstances. But the potential as well as the limits of medical research should be appreciated before considering the replication of medical research protocols to the development-evaluation domain.

In practice, peer-reviewed medical-research studies disseminated by the mass media have advertised different conclusions regarding the health benefits of such treatments as the regular intakes of vitamins, taking an aspirin a day, sleeping more than eight hours a night, drinking red wine at every meal, the cancer risks associated with using cell phones, living near a high-power transmission line, etc. Extravagant and sometimes fraudulent claims have slipped through the peer-review process of scientific journals, e.g. one large RCT found that secret prayers by unknown parties can save the lives of heart surgery patients while another proved that it can harm them (Freedman, 2010).

John P.A. Ioannidis, Director of the Prevention Research Centre at Stanford University, has designed a mathematical model for assessing the probability that a medical research finding is true (Ioannidis, 2005a). His landmark article confirms that the probability of hypotheses depends on much more than the confidence interval threshold set at 5 percent by most journals. Specifically his simulations show that the following have had a devastating effect on the validity of most published research findings:

- poor selection of the relationship being tested;
- inadequate power of statistical designs;
- medical treatments characterized by small effects;
- diverse sources of researcher prejudice, etc.

Even modest levels of researcher bias (either fed by ambition or conviction) are conducive to misinterpretation of statistical tests, distorted use of evidence and/or misleading presentation of results. Published medical research findings are often demonstrably false. Even highly acclaimed research findings can be untrustworthy (Ioannidis, 2005b).

Medical research has been captured

In part, erosion in medical-research credibility reflects changes in the institutional environment. Until the 1980s, drug research was largely independent of the pharmaceutical companies. This is no longer the case: clinical trials are now controlled by private multinational companies and RCTs do not protect the process from many systemic biases (House, 2008):

- New drugs are often tested against placebos (the selected counterfactual) rather than drugs currently in use for the same ailment. This overstates the benefits of new offerings.
- Comparisons among competing drugs are not always based on equivalent dosages.
- Younger subjects who suffer less from side effects are used for tests even though the drugs are more often targeted to older patients.
- Time scales are frequently manipulated, i.e. testing is often of short duration even for drugs taken over a life time.
- Companies, not researchers, control data analysis and publication so that findings from negative or inconclusive trials are usually suppressed and reports are written to show products in a favourable light.

Even if the research is carried out by universities, most trials are now funded by drug companies under contracts that restrict academic freedom by giving private sponsors tight control over evaluation designs, data analysis, research interpretation, dissemination of findings, etc. This is combined with a gradual and deliberate capture of the regulatory framework by private interests.

All this adds up to a serious need to promote higher-quality, independently-validated medical research. It highlights the need for more rigorous assessments of what works, and what doesn't work, through sound evaluation methods, including experimental methods. But it also throws light on the sobering fact that medical research as currently practiced is not an ideal worth striving for.

Only ethical principles and agreed standards of practice stand in the way of evaluation capture by private or partisan interests. The bottom line is that all evaluation methods (including RCTs) are vulnerable to weak priority setting in programme allocations, misleading selection of comparators, cherry picking of data, biases in reporting of findings, financial leverage, etc.

This is certainly the case in development cooperation where (with the notable exception of the multilateral development banks and a few bilateral aid agencies) evaluation functions are rarely shielded from external influence, especially at a time when pressures to show that 'aid works' have become intense.

An aid industry in turmoil

The travails of the global economy underlie increasingly adverse public attitudes towards development aid. Given unprecedented austerity measures, huge public-debt burdens and threatened living standards in western countries, rich countries' electorates have become aid sceptics and they are apt to favour evaluation methods that are (or appear to be) rigorous and scientific.

Official aid agencies are under intense public scrutiny. More than ever, they must show that they are delivering results. Accordingly there is widespread thirst among policy makers to ascertain cause–effect relationships in an unstable, noisy and risky operating environment. This helps explain the surge of interest in rigorous attribution of results to development interventions. Another likely contributing cause for the recent methodological trend is the growth of private giving for development.

In 2008, US$53 billion in aid may have been provided by civil society organizations and philanthropic sources – more than 10 times the amount earlier in the decade and close to a third of total aid flows (World Bank, 2011). The new private actors have been instrumental in putting forward alternative approaches to aid delivery and in reorienting the aid debate towards results. Since experimentalist methods focus on results, they are highly prized by private aid givers. Thus, the fundamental public rationale of experimental designs is that they promise valid results. Is this promise reliable?

The pros and cons of experimental designs

A good evaluation enjoys the capacity to persuade through compelling narratives, logical reasoning and sound methods.

The potential of experimental methods

In the right circumstances, experimental methods do establish causality by providing a valid measure of what results would have been observed had the intervention not taken place. They do so by achieving strict comparability between control and treatment groups as a result of random selection of beneficiaries and non-beneficiaries of a given intervention drawn from the same population through an explicit chance-based process (e.g. a roll of the dice; a roulette wheel; or a random number table). Unbiased allocation means that the probability of ending up in the control group or the treatment group is identical.

When expertly implemented, this approach addresses the issue of *selection bias*, which arises when comparing impacts on two very different sets of beneficiaries that may end up falsely attributing the observed results to the intervention even though different known or unknown characteristics of the treatment and non-treatment groups may have been at work.

This includes frequent cases where those who access the programme are richer, more powerful, more motivated or more educated. Random assignment to the treatment and non-treatment groups from the same population ensures that, except for chance fluctuations, the impact of the intervention can be reliably ascertained by comparing outcomes among the two groups by ensuring all the other factors that may affect outcomes are identical except for stochastic errors.

To ascertain the reliability of the finding, statistical testing techniques are available to determine the range of confidence that one may safely attribute to the result (i.e. the role that pure chance associated with the randomization process may have played). It follows that RCTs enjoy the additional advantage of allowing evaluators to establish a measure of statistical significance to evaluation findings. What then is the applicability of RCTs for assessing the impact of development interventions?

Limitations of the experimental approach

Experimental methods are not always appropriate. They are redundant when no other plausible explanation for the results observed is available. Also, they may not constitute a feasible option. For example, it is not possible to randomize the location of infrastructure projects (Ravallion, 2009). Nor are experimental methods feasible when no untreated target group can be identified; for example, when an intervention is intended to be universal (the imposition of a legal limit for alcohol consumption, a civil service reform programme, the liberalization of an import regime, etc.).

Experiments may not even be decisive in establishing attribution. This is because inferences can only be established with confidence if the treatment group and the control group and the process that affects each are strictly identical (except in terms of cause and effect). Internal validity may be jeopardized by latent and unobserved causal factors that are not taken into account when constructing the treatment and control groups.

This means that randomization is mostly suited to simple projects with easily identified participants and non-participants and where spill over effects are not likely to bias the results. Experimental 'black boxes' are poorly suited to the evaluation of complicated or complex programmes in unstable environments. Yet, this is where knowledge gaps are the deepest.

Nor is external validity the *forte* of experimental methods. Even where experiments are appropriate, they may not meet the needs of policy makers who are vitally concerned not so much with what happened in a trial experiment but with whether they are likely to keep working in a diverse, complex and volatile implementation environment (Cartwright and Munro, 2010). Programme size, structure and context matter a great deal in shaping the outcome of development activities.

Without a theory that has survived validity tests there is no credible explanation. A deep understanding of how a particular programme operates is critical and the validity of the theory on which it is predicated must be established. Securing an adequate understanding of causal relationships and identifying the rival explanations that need refutation call for substantive knowledge of the intervention, its design, its implementation protocols and the incentives of programme participants and beneficiaries.

Even where experiments make sense to assess attribution, they require superior skills, large studies, large samples and specialized quality-assurance arrangements. These prerequisites may not be available and they may not translate into an economic use of scarce evaluation resources. They may inhibit resort to cheaper and more effective evaluations. They may also hinder fulsome participation of aid recipients in the evaluation process by shifting the control of sophisticated impact evaluation to well-endowed universities and think tanks located in developed countries.

Privileging public interventions that are evaluable through experimental methods encourages the selection of simplistic programmes and projects that may not be fit for purpose and/or promote avoidance of critical evaluation questions. Most high-level policies, programmes and projects that are now privileged by international-development agencies are not evaluable through randomized treatment.

Finally, depriving members of the control group of a useful treatment based on a selection process perceived as capricious and arbitrary can be discriminatory and may even be illegal. In some jurisdictions, no comparison group is allowed to receive any treatment that is less than the best currently available. Nor is it usually considered ethical to induce members of a treatment group to participate in an intervention that may have negative side effects. Informed-consent procedures used in such cases may introduce a selection bias.

There are alternatives

Many evaluators go through their whole career without ever using an RCT. In part this is because other methods are better equipped to address issues of *why* interventions succeed; *whether* design or implementation problems explain observed intervention failures; or *who* among development partners is responsible for particular outcomes. They involve participation, observation, analysis of text-based information, village meetings, open-ended interviews, etc. Of course, qualitative-data collection requires careful coding and systematic quantification in order to be econometrically analysed.

Qualitative methods guided by theories of change examine what has actually happened and why. They are better equipped to determine the reasons for success or failure of achieving intended effects (and the extent and nature of unintended effects). In particular, they help to discriminate between design issues and implementation problems. Whereas experimental methods are shaped by data, qualitative, theory-based approaches are shaped by the questions of interest to stakeholders and the assumptions embedded in programme and project interventions (Bamberger et al., 2010).

Finally, a wide variety of tools exist to simulate a counterfactual, short of randomization. The listing that follows is only indicative of the wealth of methods and tools available to evaluators. It is not meant as an assessment of their respective strengths and weaknesses in diverse evaluation contexts.

Regression and factor analysis. Regression analysis is used to ascertain the extent to which various characteristics of the context and the beneficiaries of an intervention explain the variations in outcome effects. The balance is attributable to the programme on the assumption that all rival explanations have been factored into the model. *Regression discontinuity* compares the effects of treatment on subjects selected according to a particular criterion (e.g. expert rating of subjects on their likelihood of success or their need for the intervention). It compares the effect of the treatment just above an eligibility cut-off point with those just below.

Quasi-experimental designs. Where randomization is not feasible, one may simulate it through *quasi-experimental* designs. The individuals included in treatment and non-treatment groups are *matched* to ensure that they are similar with respect to the characteristics that may influence the outcome. Statistical adjustments are available to help ensure that the two groups closely resemble each other with respect to these relevant dimensions.

Multivariate statistical modelling. Designed to take account of all postulated relationships among treatment and non-treatment variables, the model should be capable of explaining the differences between the two groups at the initial stage so that the differences observed at the post-treatment stage can be netted out statistically. But this approach has problems of its own: it assumes that:

- the model has captured accurately the relationships among variables;
- all factors that explain the pre-treatment differences have been identified.

Participatory approaches. Qualitative impact assessment relies on the voiced perceptions of actual or potential beneficiaries, expert observers and/or decision makers. Colour voting facilitates principled debate by displaying stakeholders' opinions through coloured presentations of their votes (or scores) on clearly formulated questions about the intervention. Concept mapping involves the use of flip charts and cards (or data processing software) to obtain a graphic image of stakeholders' perceptions of the potential impacts of a development intervention. It uses skilled moderators to engage a representative group of stakeholders who are knowledgeable and committed to participate.

Surveys and sampling. Survey data collection and interpretation, structured or semi-structured interviews, focus groups and other methods of involving beneficiaries can illuminate what works, doesn't work and why. Where large groups of citizens or beneficiaries are surveyed, data collection and interpretation calls for effective sampling strategies.

General elimination methodology. Michael Scriven (2008) has proposed an alternative to RCTs inspired by criminal investigation techniques that focus on motives, means and opportunity. The general elimination methodology requires a survey of the literature and/or consultation with individuals who possess tacit expertise relevant to the intervention domain. The process starts with a systematic listing of possible causes that pertain to the intervention. Next, a list of the modus operandi for each possible cause is constructed. This is followed by a detailed examination of the facts of the case. Only the causes 'left standing' are retained as potential explanations.

Expert panels. Using expert panels of independent specialists familiar with the domain of the intervention can be useful in conjunction with other methods especially where the evaluation team does not include subject matter specialists or senior evaluators. Panels can be used to assess whether observed impacts are in line with what may be reasonably expected in a specific context. The validity and reliability of expert panels' judgements can be enhanced through a *Delphi process* that consists in consultation procedures with the individual experts without any prior consultation among them.

Benchmarking. Internal benchmarking identifies and seeks to replicate good practices observed within a programme. External benchmarking compares the impact of an intervention with that of a similarly situated initiative perceived to have achieved standards of excellence. Benchmarking uses key performance tests to judge impact through comparisons with good or best practice observed in similar circumstances.

Perhaps the only valid generalization that can be gleaned from the above is that evaluation tools including RCTs are just tools. They should not be allowed to dominate what is first and foremost a creative, analytical and participatory process. Experimental methods have many powerful statistical features that other evaluation designs cannot easily match in some circumstances. But a threat to good evaluation management is overinvestment in a single technique. A tool can only fulfil the function or functions that it was designed for.

Wielding the right tools

Of course using the right tools and using them with care and skill is an important ingredient of evaluation quality. Inappropriate methods can sink an evaluation. But threats to the rigour of an

evaluation may also result from other factors: sloppy data collection, politically naive evaluations; lack of independence; inadequate evaluators' competencies; failure to focus on utilization; ignoring the context; limited involvement of stakeholders; concentration on unimportant or irrelevant issues, etc.

Well-selected evaluation tools used according to their specifications contribute to the validity of evaluations. They make evaluations easier to compare and facilitate their costing and planning. When they are used judiciously, they make evaluation findings more credible and predictable. Understanding and, where possible, measuring the limits of the tools used in a particular context is critical to quality. The inability to connect the detailed design of the evaluation to the priority questions identified at the planning stage explains why many evaluations go bad.

Consequently a good understanding of the respective strengths, weaknesses and limitations of evaluation methods and tools is a critical competency for evaluators. While experimental and quasi-experimental methods can in some circumstances illuminate attribution of observed outcomes, theory-based and process evaluations are better equipped to answer how and why the observed effects have materialized. It is therefore fortunate that all national and regional evaluation guidelines and standards give adequate weight and credence to qualitative approaches: they stress methodological appropriateness rather than doctrinal orthodoxy.

Prospects

It should be clear by now that the rise of experimental methods in development evaluation is related to several factors: (i) rising pessimism about development assistance and the intense pressure on policy makers to demonstrate that 'aid works'; (ii) a mistaken faith in the medical research model; (iii) poor quality assurance in development evaluation.

Will the challenge that the randomization movement has posed for the development-evaluation profession elicit a healthy and principled response within the development community? The future is already here. Improving the rigour of evaluation work in development is well underway through increased use of mixed methods, enhanced access to evaluation training and a growing preoccupation with evaluators' capabilities. Evaluation approaches will continue to evolve and improve in line with changes in the development-cooperation enterprise.

Evolving development-assistance paradigms

The Millennium Development Goals have displaced economic growth as the dominant objective of development. This shift in emphasis is consistent with the growing recognition that national income is an unsatisfactory indicator of economic and social progress. It fails to capture highly valuable services provided within the household. It does not measure the environmental losses, the income inequalities or the social disruptions associated with unbridled economic growth. It is quality growth (and not economic growth per se) that constitutes the overarching economic, political and ethical imperative of the contemporary development enterprise.

In the wake of an unprecedented financial crisis that has turned decades of economic orthodoxy on its head, development thinking has come to a crossroads. Amartya Sen, the Nobel laureate, has equated development with freedom (Sen, 1999). In the same vein, the three dimensional (3D) model proposed by Allister McGregor and Andy Summer (2010) offers a timely analytical tool that captures the material, relational and perceptual characteristics of human aspirations and social progress and it provides a convenient framework for assessing philanthropic development interventions whether focused on improved capabilities or more favourable enabling conditions.

Table 1. Human well-being and evaluation

Evaluation characteristics	Material well-being	Relational well-being	Perceptual well-being
Major discipline	Economics	Sociology	Psychology
Dominant evaluation approach	Cost–benefit analysis	Participatory evaluation	Empowerment evaluation
Investment focus	Physical capital	Social capital	Human capital
Main unit of account	Countries	Communities	Individuals
Main types of indicators	Socio-economic	Resilience	Quality of life

Evaluation methodologies and practices have always adapted to reflect shifts in policy paradigms and to serve evolving social policies. Beyond the parsimonious approaches associated with experimental methods, the matrix below points to the pluralistic methods that will have to be marshalled to do justice to the holistic conception of development embedded in human well-being aspirations.

What does the future hold?

The current surge of interest in randomization reflects a reaction to the lax quality standards and the positive bias of many development evaluations especially those commissioned and controlled by programme managers. But experimental methods have a strictly limited role. In principle they are the instrument of choice to assess attribution but in practice they are appropriate only for relatively simple interventions, the effects of which are realized in a short period of time and are large relative to other potential influences. Even where they are suitable, experiments are exceedingly hard to implement rigorously given their high costs, contamination risks and exacting statistical and ethical requirements.

Frequent claims that the bio-medical clinical trial procedure holds the key to evaluation rigour in the economic and social domain are invalid. They are rooted in naive perceptions of the actual record and they do not acknowledge that the issues that experimental methods can address are constrained by the limitations inherent in the method or that they are inappropriate for complex and adaptable programmes implemented in diverse settings. Other methods than the experiment are available to answer the wide range of questions policy makers would like answered.

A widespread yearning for combining accountability with learning in development cooperation underlies the current experimentalist craze. This yearning can only be satisfied by independent and self-evaluation functions that draw on the full panoply of evaluation methods. Effective triangulation approaches resort to experimental methods only where feasible and appropriate. They use qualitative techniques, participatory methods, beneficiary surveys and carefully constructed case studies.

These methods are more 'fit for purpose' in the development domain than RCTs (Bamberger et al., 2009). Looking ahead, experimental methods will be used for a limited range of interventions that have been specifically designed to make them feasible. More interventions will be conceived as policy experiments but they will be assessed through mixed methods. The current bubble of enthusiasm for experimentalism in development evaluation is bound to burst.

Acknowledgment

This viewpoint article is based on a presentation delivered at the 9th European Week of Regions and Cities (Open Days) during a workshop *'How to capture the effects of EU funding? Bringing together qualitative and*

quantitative methods' chaired by Kai Stryczynski and co-sponsored by the European Evaluation Society on 11 October 2011.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Notes

1. A welcome exception is the special issue of *Review of World Economics*143(4), 2007 about various aspects of aid heterogeneity that led the editors (George Mavrotas and Peter Nunnenkamp) to conclude that 'substantial effort needs to be taken to delve deeper into the various routes and transmission mechanisms through which the various types of aid operate'.
2. http://www.williameasterly.org/
3. In medicine, a gold standard test refers to a diagnostic test or benchmark that is regarded as definitive.
4. The intra-evaluation methodological conflict flared again briefly in late 2003 in the USA when the Department of Education ruled that experimental methods would be privileged in its evaluation funding.
5. http://www.worldbank.org/ieg/nonie/about.html
6. J.S. Mill's most significant works include *A System of Logic*, *Principles of Political Economy*, *On Liberty*, *Utilitarianism*, *The Subjection of Women*, *Three Essays on Religion*, and an *Autobiography*.
7. Inspired by Comte, J.S. Mill also adhered to a Religion of Humanity – an atheist doctrine that idealizes and reveres humanity while embracing the core ethical features of traditional religion.
8. Voltaire (1694–1778), Rousseau (1712–78), Montesquieu (1689–1755), Buffon (1707–88), Turgot (1727–81), David Hume (1711-1776), Adam Smith (1723–90), Edward Gibbon (1737–94) and Immanuel Kant (1724–1804) among others.
9. The advent of the evaluation discipline also coincides with the origins of the development enterprise – a time of optimism when the swords of the Second World War were turned into ploughshares by the victorious allies.

References

Alkin MC (2004) *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: SAGE. illshiills

Bamberger M, Rao V and Woolcock M (2009) Using mixed methods in monitoring and evaluation: experiences from international development. Brooks World Poverty Institute Working Paper 107, University of Manchester (www.manchester.ac.uk /bwpi).

Campbell DT (1974) Qualitative knowing in action research. Kurt Lewin Award address, Society for the Psychological Study of Social Issues, presented at the meeting of the American Psychological Association, New Orleans, LA, September.

Campbell DT and Stanley JC (1963) *Experimental and Quasi-Experimental Designs for Research.* Chicago, IL: Rand-McNally.

Cartwright N and Munro E (2010) The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice* 16(2): 260–6.

Center for Global Development (2006) When will we ever learn? Improving lives through impact evaluation. Report of the Evaluation Gap Working Group, Washington, DC.

Cronbach L (1982) *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey Bass.

Deaton A (2005) Some remarks on randomization, econometrics and data. In: Pitman GK, Feinstein ON and Ingram GK (eds) *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, Volume 7. New Brunswick, NJ and London: Transaction Publishers.

Duflo E and Kremer M (2005) Use of randomization in the evaluation of development effectiveness. In: Pitman GK, Feinstein ON and Ingram GK (eds) *Evaluating Development Effectiveness*, World Bank

Series on Evaluation and Development, Volume 7. . New Brunswick, NJ and London: Transaction Publishers.
Freedman DH (2010) Lies, damned lies and medical science. *The Atlantic*, November.
House ER (2008) Blowback: consequences of evaluation for evaluation. *American Journal for Evaluation* 29(4): 416–26.
Ioannidis JPA (2005a) Why most published research findings are false. *PLoS Medicine* 2(8): e124.
Ioannidis JPA (2005b) Contradicted and initially stronger effects in highly cited clinical research. *Journal of American Medical Association* 294(2): 218–28.
Jerve AM and Villanger E (2008) *The Challenge of Assessing Aid Impact: A review of Norwegian Evaluation Practice,* Norad, Study 1.
Leeuw F and Vaessen J (2009) *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*. Washington, DC: World Bank.
McGregor A and Sumner A (2010) Beyond business as usual: what might 3-d wellbeing contribute to MDG momentum? *IDS Bulletin* 41(1): 104–12.
Picciotto J (2010) *Labors of Innocence*. Cambridge, MA: Harvard University Press.
Ravallion M (2005) Comments. In: Pitman GK, Feinstein ON and Ingram GK (eds) *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, Volume 7. New Brunswick, NJ and London: Transaction Publishers.
Ravallion M (2009) Should the Randomistas Rule? *Economists Voice* 6(2): article 6.
Rosenau PM (1991) *Post Modernism and the Social Sciences: Insights, Inroads and Intrusions*. Princeton, NJ: Princeton University Press.
Sachs J (2005) *The End of Poverty: Economic Possibilities for Our Time*. New York: Penguin.
Scriven M (2008) A summative evaluation of RCT methodology and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation* 5(9): 11–24.
Sen A (1999) *Development as Freedom*. Oxford: Oxford University Press.
Tarp F (2009) Aid effectiveness. United Nations University, WIDER, Helsinki. URL: http://www.un.org/en/ecosoc/newfunct/pdf/aid_effectiveness-finn_tarp.pdf
White H (2009) Some reflections on current debates in impact evaluation. International Initiative for Impact Evaluation, Working Paper 1, New Delhi.
World Bank (2011) *Global Monitoring Report 2011: Improving the Odds of Achieving the MDGs*. Washington, DC: World Bank.

Robert Picciotto, AcSS, is Visiting Professor, King's College, London. A former Director General of the Independent Evaluation Group of the World Bank, he currently serves as board member of the European Evaluation Society and as council member of the UK Evaluation Society. Please address correspondence to: King's College London, Strand, WC2R 2LS, UK. [email: r.picciotto@btinternet.com]