Imagining the future of evaluation

Anne Revillard, Tom Cook, Sandra Mathison, Rebecca Maynard, Ray Pawson, Laura R Peck

Débats du LIEPP

SciencesPo
#9
février 2025

Imagining the future of evaluation

Introduction

Anne REVILLARD Sciences Po

Getting back to "what works"?

Tom COOK Northwestern University

Evaluation for the public good

Sandra MATHISON University of British Columbia

Keeping the end-users in mind

Rebecca MAYNARD Penn Graduate School of Education

Broadening the focus

Ray PAWSON University of Leeds

Towards an evaluation culture

Laura R. PECK MEF Associates

How to cite this publication:
Anne Revillard, Tom Cook, Sandra Mathison et al., **Imagining the future of evaluation**, *Débats du LIEPP n°9,* 2025-02-03.

Introduction

by Anne Revillard

Based on a seminar organized by LIEPP which took place at Sciences Po in May 2024, this publication brings together contributions from five leading scholars in evaluation to imagine the future of the field. What new methods and approaches should be promoted? How does evaluation differ from applied social science methods – and should it be any different? What are the major institutional and political challenges to the design and conduct of evaluation and the promotion of its use in policymaking and within civil society? What should be the role of evaluation in democracy, and what practical tools can it rely on to fulfill this role? These are some of the questions addressed in this "LIEPP Debates". To this end, LIEPP has gathered researchers with very different theoretical and methodological orientations in evaluation.

Tom D. Cook is Professor Emeritus of Sociology at Northwestern University. He is, with Donald Campbell, one of the early promoters of experimental and quasi-experimental evaluation and he has played a major role in the development of evaluation theories. Showing how evaluation has drifted away from its foundational "what works?" question to include reflections on items such as implementation, program theory, utilization or mixed methods, Tom Cook argues we should re-centre evaluation theory and practice around impact measurement ("what works?"), since establishing whether the program has any effect should be preliminary to all these other evaluative questions.

Sandra Mathison is Professor of Education at the University of British Columbia. She has significantly contributed to the theory and practice of qualitative methods in evaluation as well as participatory approaches. She is a major critical and transformative voice regarding the role of evaluation in neoliberal societies. In her contribution, Sandra Mathison raises the question of evaluation's relation to democracy and the extent to which evaluation can contribute to the public good by "speaking truth to the powerless". Acknowledging trends such as the rise of data-driven decision-making, real-time evaluation, and the focus on equity and diversity, she notably insists on the need to engage in macro-analysis of evaluation theory and practice, to always keep in mind the question of relevance ("does this program even ought to exist, let alone be evaluated?"), and in view of this, to promote independent evaluation.

Rebecca A. Maynard is Professor Emerita of Education and Social Policy at the Penn Graduate School of Education. She is a leading expert in the design and conduct of randomized control trials and rapid cycle evaluation in education and social policy. She has also significantly contributed to the development of open science practices in evaluation in terms of pre-registration, data sharing and evidence synthesis. Taking stock of some assets in contemporary evaluation practice (a diversity of methodological skills, strong training programs, more opportunities for interdisciplinary work, the development of evidence repositories), Rebecca A. Maynard argues we can make evaluation more useful by keeping the end-users in mind at all stages of evaluations, by grounding more systematically the evaluation design in the existing knowledge base, by adapting the methods to the study goals, the context of the theory of change, and by making research more accessible.

Ray Pawson is Emeritus Professor of social research methodology at the University of Leeds. His distinct contribution to evaluation theory and practice draws on critical realism with the development of realist evaluation as an approach that addresses complexity and replaces the classical evaluative question of "what works?" with the more complex and specific question of "how does it work, for whom, in which circumstances?" He draws on his latest book, *How to Think Like a Realist*?[1], to identify several suggestions for the future of evaluation. He stresses

1 Pawson, Ray. *How to Think Like a Realist. A Methodology for Social Science*. London: Edward Elgar Publishing, 2024.

the need to go beyond a commissioner-driven "atomised inquiry" practice of evaluation and to better integrate evaluation findings to promote cumulative learning. He also advocates in favour of funding more *ex-ante* evaluations, broadening the focus from programs to a broader array of policy instruments, and integrating historical analysis into evaluation.

Laura R. Peck is principal scientist and director of the Income Security and Economic Mobility domain at MEF Associates. She was previously a tenured associate professor at Arizona State University. She specializes in experimental and quasi-experimental evaluation to which she has made major contributions, notably through her reflection on integrating some of the critiques leveled by the U.S. evaluation field regarding the need to open the black box of interventions, which her book *Experimental Evaluation for Program Improvement*[2] (2020; SAGE) explores. In her contribution to this LIEPP debate, she insists on data quality, the need to embed equity principles into evaluation, and fostering learning agendas to promote and evaluation culture, in order to go beyond the current "feedback/satisfaction/surveillance culture".

Taken together, these contributions illustrate the many methodological, practical and political challenges the field of evaluation currently faces, as well as its assets for the future.

[2] Peck, Laura R. (2020). *Experimental Evaluation Design for Program Improvement.* Thousand Oaks, CA: SAGE Publications.

Getting back to "what works"?

by Tom Cook

The past is said to be the best predictor of the future, and so I will introduce you to my version of evaluation's history. I do this to examine the future I think will occur in Evaluation while also regretting a desirable future that is not likely to occur. To describe my position, I will briefly introduce what I call the centrifugal theory of the history of evaluation. It illustrates how evaluation theory and practice have evolved over the last 50 years or so and describes how this evolution has weakened links to learning both "which programs work" and "how to improve effective programs". Today, I contend that Evaluation is best characterized as use of all social science theories and methods applied to whatever social issues are of interest to stakeholders who occupy (often peripheral) policy positions. Sometimes, these stakeholders want to learn about program effectiveness, but I contend that this is comparatively rarely. We ask: Is Evaluation increasingly concerned with issues that have an unclear or minor relationship to assigning value to social programs, even though assigning value is a key task of Evaluation, if not its main one?

At its beginning, evaluation was dominated by two academic fields. One was labor economics, a field that has historically been concerned with learning what works to improve labor markets and individual welfare within them. Now, it deals with substantive issues beyond labor and might be better described as applied micro-economics. The second field was Psychology, especially social and educational psychology and the prevention sciences, where researchers were primarily trained in a laboratory experimentation tradition that privileged descriptive causal questions of the type: Does manipulating variation in A cause a change in B? The earliest modern evaluations reflected such a framing, asking: Does an existing program (or plan for a program) work to bring about positive benefits? The preferred methods for answering such a question were broadly experimental.

However, within a decade the results from this framing proved to be disappointing. Most evaluations showed no demonstrable effects; and if they did, the impacts were disappointingly small or even seemed, in the case of an early Head Start evaluation, to be negative. Moreover, when researchers made claims about effects their claims were not treated as nuggets of truth from on high; instead, they were almost invariably disputed either on technical grounds or because critics thought there might be sub-populations of interventions, outcomes, persons or situations where the effect would hold but had not been included into the study sampling design or had not been properly analyzed for. Also obvious was that policy makers were not waiting eagerly for the evaluation results and, if they did get them, did not do anything to apply them. All this set off a series of centrifugal forces that quickly spun evaluation away from its simple question -- Does the program, policy or practice work?

Yet the earliest reaction to the initial disappointment was centripetal rather than centrifugal. Some researchers doubled down on a circumscribed version of the original experimental agenda, rejecting quasi-experimental methods and seeking to limit Evaluation to experiments with random assignment. The claim was that quasi-experiments produced biased results because of a pervasive "selection bias" that predominantly operates to underestimate effects, sometimes completely obfuscating them. So, randomized experiments became heavily preferred for developing causal answers, their warrant coming from basic statistical theory. But many important policy-relevant causal questions do not lend themselves to random assignment, particularly questions about large-scale and universal programs. Moreover, experiments are limited for answering questions about the contingencies determining when a program has meaningful effects and for answering explanatory questions of the type: Why (or how) does a given program generate its

effects? Random assignment answers the more modest question of whether A as manipulated causes change in B as measured. The randomized experiment lost its privileged status for these reasons and also because it did not lead to exciting findings about effective programs, policy makers outside of medicine and health did not particularly privilege it, and as Evaluation grew as a sub-field with resources for research studies, researchers from other fields with different method preferences entered the field from Sociology, Political Science, Public Policy, Anthropology and even Philosophy. All this novelty spun evaluation away from identifying "what works" and "improving programs that work" towards "using any kind of method for any task in applied social science writ large".

Foremost among the earlier centrifugal forces was realization that the policy use of evaluation results is complex and problematic and, in the social sciences at least, is rarely characterized by policy makers taking the results of single studies and using them to structure their decision-making. Policy choices are the product of many forces, of which rationally generated information is only one and rarely a dominant one. So, a concern with utilization developed and became an area of study within Evaluation called "*utilization-focused evaluation*", with the adjective reminding readers that evaluation topics should be chosen for their high likelihood of immediate policy use and not because social programs, or ideas for programs, were involved that might eventually lead to important social change. Concern with use is also central in "*practical evaluation*", where the emphasis is on researchers making themselves available to policy makers to generate whatever knowledge they think they need that might, or might not, include causal knowledge. After all, program plans, surveys, small-scale interviews, or advice about implementation might be more useful to policy-makers in general than knowledge of what works. Indeed, some policy actors may even seek to subvert outcome-oriented evaluation for fear its results will be used to hold them accountable for the programs they manage. To disburse evaluation funds for evaluability assessments, surveys, or on-site observations does "evaluation" but in ways that are generally seen as irrelevant to effectiveness.

Another centrifugal force in evaluation's early history was a concern with how social programs are implemented, since inadequate implementation might explain the paucity of early positive results. When planning an evaluation, it is often difficult to specify the implementation levels and processes needed for effectiveness, just as it is difficult to predict all the local complexities that can arise when mounting an intervention. Systematic implementation studies arose in Political Science for monitoring the Great Society programs in the USA, and the concern was to examine implementation in its own right without special consideration of its links to "what works" other than through the obvious beliefs that (a) a poorly implemented program almost certainly cannot work and (b) identifying implementation shortfalls can help improve a program that has already been demonstrated to work. However, learning how to improve implementation will not improve benefits if a program is ineffective. So, why evaluate the implementation of a program whose theoretical or empirical fundamentals have not been shown to be valid and that might never be succeed under any realistic scenarios? Implementation quality is frequently a problem in Evaluation, and its study has a legitimate place within the field. But the issues are: How to ensure that implementation activities are not wasted on programs that do not work; and how to prioritize implementation and causal issues within the same project, since adding more evaluation goals with a finite research budget will likely weaken the quality of evidence generated about any one goal.

A third centrifugal force emphasizes the study of "*program theory*". Program theories try to specify (a) all the main forces the sequentially follow after introducing an intervention and before measuring its major planned outcomes, and (b) all the human, material and financial resources needed to move from one node to another in the postulated chain of causal

relationships. Specifying such throughputs is useful for learning why "what works" works; and it highlights the benefits of testing early outcomes as well as the final ones on which experiments tend to prioritize but that are further removed from the intervention and more difficult to realize in practice. Program theories have the further advantage of sometimes pointing to the kinds of persons, settings and program variants that an intervention is more likely to affect. However, they have two major problems. First, an emphasis on program theory is convenient for policy actors wishing to escape accountability; they can sponsor a pre-evaluation exercise that postpones or avoids the kind of outcome evaluation whose disappointing results might threaten their interests. And second, while planning can often point to what is not likely to work, it cannot by itself be informative about program effectiveness. At best, it suggests that a program could work if the theory behind it were true and if all the intervening steps it postulates are realized on the ground. Evaluating program theories does not assign empirical value to a program, though assigning such value is the functional root of evaluation as a concept.

A fourth centrifugal force is the advocacy of "*realist evaluation*". This rightly distinguishes between different theories of cause, treating causal contingency and causal explanation as ontologically real and more valuable than experimental causal descriptions because they are more likely to advance program theory and to improve program design. They do this (a) by identifying the causal contingencies determining when and for whom a program is effective under the assumption that the error terms over which effects are generalized in most quantitative analyses cannot identify such contingencies; and (b) by identifying the causal explanations that describe how or why a program works so that programs can be designed or improved for circumstances different from those examined to date. These two conceptions of cause are indeed important and re-state the current scientific consensus about the relative value of different understandings of cause. But there are significant problems too. One is the high likelihood of not achieving either goal to a level even close to closure about the knowledge gained. While advocates of realist evaluation are willing to use any quantitative or qualitative method to their ends, in specific evaluation projects the contingency approach never details any more than a few of the many causal contingencies, while the explanatory approach rarely identifies which of the many possible (or even plausible) causal pathways possible is the valid one. The ontological assumptions are realist, but the method choices are not realistic in current practice without lowering standards of causal evidence and letting researcher preferences inadvertently slip into the conclusions drawn. Another problem is that identifying causal contingencies and mechanisms pre-supposes an effective program. If a program is independently known to be helpful, then the further study of contingency and explanation is warranted and likely to be helpful to program improvement; but if it is not known to be helpful, then single studies will be called upon to establish a causal relationship, to identify its contingencies, and to adequately test among plausible theories of its causal mechanism. All this loads up a study and, in practice, requires prioritizing among these different goals and the resources devoted to each. How these priorities are structured will have important consequences for the accuracy of the knowledge gained about any one goal, and it is not easy (for me at least) to see any main priority other than "what works" since this question is logically prior to probing causal contingency and causal explanation in a resource-responsible way. Why probe contingency and explanation for a program with no overall effects or not yet known to have any? And even when there are no overall effects, some sub-groups might indeed show a positive effect, but a serious ethical conundrum then arises because it logically follows that the program will have actively harmed other groups.

The final centrifugal force we examine is *mixed method evaluation*. Social science methods have evolved to provide answers to different kinds of question, many of which are non-causal in any usual sense of that word. Given the link between methods and purposes, the call for mixed method evaluation is essentially a call for multi-purpose evaluation, for consideration of the theory, planning, implementation, causal-description,

causal-explanation and utilization purposes documented earlier. At the practice level, though, the advocacy of mixed methods is often framed around incorporating some quantitative and some qualitative methods into the same evaluation project to answer some among the many different kinds of questions with which Evaluation now deals. (The call for mixed methods also reflects researchers from different disciplines entering into Evaluation and bringing with them their discipline's particular substantive issues and method preferences). Mixed-method evaluation is not needed in evaluation projects that address a circumscribed question, for a causal-descriptive question might require just an experiment, or an implementation issue might require on-site monitoring without any quantitative analysis. Of course, the conclusions from such a circumscribed study might lead to subsequent questions and these in their turn might require yet other methods. That is fine, since most multi-study programs of research into the same social program will surface quite different kinds of questions. A second limitation is the absence of extended, grounded analyses of how to conceptualize and act on the inevitable trade-offs among both questions and methods. Resource constraints will force evaluators to accept methods for answering some questions that are far from technically optimal, and in contexts where effectiveness is not already known it will be difficult to prioritize on issues other than effectiveness, for causal contingency and explanation logically depend on having already demonstrated what works.

Let me state the conclusions of this talk as a few propositions.

1. The history of Evaluation as a field of study has evolved from a narrow focus on experimental methods for inferring program effectiveness to a field that applies all social science methods to examining whatever issues policy makers raise. These can touch on program theory, evaluation planning, program implementation, identifying program effectiveness, specifying causal contingencies, uncovering causal explanatory mechanisms, and learning about how to promote the utilization of evaluation results. Since these different knowledge needs require different methods, Evaluation currently espouses multi-method research practices.

2. Broadening the range of Evaluation's issues and questions is very useful because effective programs can generally be improved the more is known about their theory and implementation and about the causal contingencies and explanations that influence the level of obtained effectiveness.

3. But evaluation tasks other than testing program effectiveness can also be irrelevant in some real contexts, as when (a) they are used to postpone evaluations to establish effectiveness, (b) their knowledge gains about theory, implementation and utilization shed little or no empirical light on whether a program works, (c) their probes of causal contingency and explanation are for programs whose effectiveness is not demonstrated well or that can never be effective, and (d) their evaluation results might be used harmfully because they fail to meet high standards of evidence.

4. Is a future task to make sure that more of the activities carried out as Evaluation are directly linked to improving knowledge of what works? Otherwise, Evaluation will continue to be a field that uses a wide array of social science methods to respond to the knowledge needs of policy-makers, many of whom deal with matters of limited consequence for human welfare despite the need to discover what works and how to make it work better. The reality of evaluation practice as I have sketched it above may well predict its immediate future. But can an alternative future be realized that grounds the range of current evaluation practices more tightly into identifying and improving what works? Otherwise, Evaluation is less useful in its societal yield and reduces itself to a minor sub-field within applied social research.

Evaluation for the public good

by Sandra Mathison

As I think about the future of evaluation, I am of a mind to look back at what I expected evaluation theory and practice to be over the 40 years or more I have been working in the field. To answer the question, I want to mention some touchstones in my own work that help me to imagine what the future of evaluation might be. In 1997, I wrote a paper entitled « The Ameliorative Assumption in Evaluation[3] » in which I asserted that all evaluations, regardless of approach taken, fundamentally mean to be helpful. Of course, how making things better happens varies significantly depending on one's paradigm and one's approach to doing evaluation.

I identified two main orientations: progress through science and progress through democratic processes. Methodologically, progress through science might equate most closely with experimental and quasi-experimental, theory-driven, and systems analysis evaluation orientations, and progress through democratic processes is more associated with participatory, collaborative, emancipatory evaluation orientations. Theoretically, the former being aligned with a post-positivist perspective and an objectivist epistemology, and the latter aligned with an interpretivist perspective and a social constructivist epistemology. I think that the basic thesis of that paper still holds and that evaluators still mean to be helpful and make things better as a consequence of the work they do.

In the 2000s, my work focused on looking at how potentially positive connections could be made between evaluation and democracy and democratic principles. I focused on deliberative democratic evaluation informed by the work of Ernie House and Kenneth Howe, but my work was even more radically focused on engaging stakeholder groups in ways that were absolutely equitable and fair. It often surfaced and maybe even created conflict among groups of stakeholders, especially if there was serious differential access to power and knowledge amongst those stakeholder groups. Sometimes it led to a more democratic approach to understanding programs, thinking about whether they were working and thinking about how to make them better; and sometimes it did not work.

In the late 2000s, we saw the global demise of democracy. This global change is not the fault of evaluation, but this fact raised questions about the role and efficacy of evaluation in promoting democratic principles and governance. I began to ask myself whether evaluation does contribute to social good. Does it live up to that ameliorative assumption? To answer this question, I first analyzed the constraints of evaluation's potential [4] by illustrating several features of evaluation that get in the way of its role in promoting democracy and equity: for example, defining social problems and solutions primarily in market terms, the commodification of evaluation itself, the co-optation and complicity of evaluation and evaluators in the neoliberal nexus of capitalism and governments.

The following year, I gave a keynote address at the Australasian Evaluation Society titled *Does Evaluation Contribute to the Public Good*[5]*?* My conclusion in that keynote address was that evaluation doesn't seem to be contributing substantially to the social good in

3 Mathison, S. "Understanding the Ameliorative Assumption in Evaluation." annual meeting of the American Evaluation Association, San Diego, Calif. 1997.

4 Mathison, S. (2016). Confronting capitalism: Evaluation for social equity. In S. Donaldson & R. Picciotto (Eds.*). Evaluation for an equitable society*. Information Age Publishing.

5 Mathison, S. (2018). Does evaluation contribute to the public good? *Evaluation*, 24(1), 113-119. https://doi.org/10.1177/1356389017749278

either a local or a global level. I'm not sanguine about the contributions that evaluation does and maybe can make. What I see is a world where actions are often not driven by evaluation or data, but rather by ideologies, and above all by neoliberalism. Of course, evaluation alone can't remedy all human problems, but we make claims that we can support the amelioration of human suffering and so we need to hold ourselves accountable, at least at some level.

So, what does the future of evaluation look like? In some ways there is an obvious future of evaluation because it builds on what's already happening in terms of approaches, methods, things that people are working on.

I think we can expect continued emphasis on data-driven decision-making. This is a long-standing orientation in evaluation, even though there's a rich literature that queries the lack of use of evaluation information. Nonetheless, this will remain an aspiration, and this will reinforce a continued emphasis on outcomes and impact.

This aspiration will also be fuelled by the availability of technology and advanced analytics, certainly characterized by technology, such as the use of apps, the skill we get with online survey capabilities, data dashboards and visualization, uses of spatial mapping, as well as machine learning, AI, predictive modelling, and all the advances that will continue to come in statistical analysis of complex data. So maybe we'll see more Bayesian methods, maybe more multi-level modelling.

At the same time, and likely in parallel, there will be continued emphasis on evaluation that is real-time and continuous. Approaches that emphasize process as much as impact and outcome but are meant to provide quick feedback in specific contexts. And this will likely continue to support the collection and use of qualitative data, sometimes instead of quantitative data, but often as a more comprehensive and complex way of understanding programs, what they are and the impact they have.

Moreover, we are likely to see continued and increased emphasis on equity and diversity, both as an orienting perspective such as in social justice-oriented evaluation, as well as engaging the diversity of stakeholders in ways that promote respect, authenticity, and differential needs. We will see evaluations continue to adopt and build on the notions of cultural responsiveness, feminism, equity, invisibility, and power relations. Evaluators will continue to use strategies that disaggregate data by factors such as race, ethnicity, gender, social class to identify and correct disparities and inequities.

Taking heed of Paul Taylor's quote about possible futures for evaluation: "We must decide what ought to be the case. We cannot discover what ought to be the case by investigating what is the case". Some thinking about a future for evaluation that steps outside of what we've been doing, even if better, is needed and here are a couple of thoughts.

One, we need to engage in more macro-analysis of evaluation theory and practice. This requires a sociological and political analysis of evaluation, less research on technical competence of evaluation, and more research OF evaluation, which leads to thinking about the moral purpose of evaluation in serving the social good. We need to ask questions about whose interests are served, whose interests should be served, and about the relationships among evaluation theory, practice, and ideologies.

Evaluation needs to not only ask whether the job is being done right or well, but also to ask if the job is worth doing. As a commodity, evaluation is purchased within a neoliberal context, which means there is little incentive to ask serious questions about whether this is a program that even ought to exist, let alone be evaluated. Evaluators also need to pay attention to what I would call the velvet hand of philanthrocapitalism and corporate social responsibility, which are merely palatable and softer versions of neoliberal capitalism.

We need to find new governance structures that allow us to think about and do evaluation in a much broader way: new forms of funding, new forms of governance, independently funded evaluations. We need to foster independent structures that can support multilateral evaluations on crucial topics that matter to the quality of life across the world, things like climate change, banking, prisons, schooling.

Mixed methods in evaluation need to be more meaningful than simply adding some qualitative data to a post-positivistically oriented evaluation. Pushing the boundaries of "mixed methods" at an epistemological level could contribute to a more fulsome and complex understanding of what is, and whether it is good or right. Earlier, I contrasted two dominant epsitemologies (post-positivist and constructivist), but there are others: non-Western ways of knowing and thinking, indigenous ways of knowing and thinking. Evaluation needs to open itself to those other possibilities.

Keeping the end-users in mind

by Rebecca Maynard

The very purpose of policy evaluation is to benefit the world—to improve the health, social and economic well-being of individuals and society – and to serve those goals through product and policy development and through effective deployment of programs, policies, and products by practitioners. However, the evidence needs of product developers, policy makers, and practitioners vary. **Product developers** need grounded theory informed by case studies and, quite importantly, they need continuous feedback to inform continuous improvement and to manage shifting needs and contexts. Policy developers and practitioners are much more focused on point in time evidence to guide their work. **Policy makers** typically need nuanced information to help them identify practical strategies for achieving some goal—what might work, under what conditions, at what cost and with what risks? **Practitioners** have similar needs but with a narrower focus—what will work to address their specific needs?

Although evaluation often is less influential than we hope, we have developed a strong and rapidly strengthening evaluation infrastructure:

1. We have rounded the corner in having an evaluation workforce with **strong quantitative, qualitative and policy analysis skills**

2. We have many very **strong graduate training** programs and a **growing set of guidelines and tools** to support evaluators' production of evidence that is useful to policy makers and practitioners

3. We have many **more opportunities for interdisciplinary work**, facilitated by open access on-line training and a growing culture of collaboration, facilitated by more open access to publication and data.

4. More **funders are supporting** development of **gap maps**, the creation and maintenance of **evidence review platforms**, and incorporation of **continuous improvement** efforts as part of program operation plans

Still, there are weak links in the system. While we have a strong and continuously improving evaluation workforce, we are much **less successful in routinely creating interdisciplinary evaluation teams**—for example, integrated teams of quantitative and qualitative researchers. Even teams with strong technical training **too often produce poorly designed and/or executed evaluations**—sometimes due to lack of funding, but more often due to inattention to important contextual nuances. Then, too often evaluators fall short communicating the findings of their evaluations. Evidence review platforms are extremely helpful in identifying evidence. But **too often the evidence bases are skimpy** or quirky in ways that diminish their usefulness.

Cost and benefit-cost analysis is critical for most policy decisions. But there is relatively **little attention to measuring costs and benefits in ways that facilitate meaningful comparisons**. For example, cost-effectiveness estimates based on standardized mean differences are not useful for policy purposes; one needs consistent unit-measures of the outcomes (e.g., years of school completed or score on the ABC test).

The public policy evaluation ecosystem is complex. There are many weaknesses, but we also have many, if not all, of the pieces to improve the usefulness and use of evidence: scoping reviews and gap maps to reveal what we know and where work is needed, realist reviews and theories of action, systematic reviews and meta-analysis, implementation and

process evaluations, rapid cycle evaluations, cost and cost-effectiveness studies.

What we lack is a strong culture of routinely knitting these pieces together. This is due to limited collaboration among evaluators from different disciplines and traditions, as well as the complexity of the issues we address, and the competing and often shifting interests and evidence needs of stakeholders among groups and over time.

This leaves us with several rooms for improvement. Here are some low-lift ways we can make our evaluations more useful and used.

1. Design evidence-building agendas with the end-users in mind. Know your end-users – the primary end user may not be the evaluation funder. consider both who an evaluation must benefit, as well as who else may benefit.

2. Ground the evaluation design in the existing knowledge base. We often do a poor job grounding future work in what we already know, especially in the case of cross-disciplinary issues.

3. Use designs and measurement strategies that are aligned with the study goals, the context and the theory of change. Too often, we have the right tool and the wrong question or vice versa. This is where interdisciplinarity can help us. Economists and sociologists now work better together in ways that elevate the usefulness of their research and evaluation.

4. Finally, we need to make research accessible and improve the usability and use of evidence repositories, especially for non-academic users and people who are going to implement practices and policies.

Going back to our eco-system, the path to improving the usefulness and use of evidence is through connecting the parts.

We should build our evaluation agendas around the needs of the end users—and we would do well to proactively enlarge the current honeycomb of knowledge. We should draw from extant knowledge when designing studies and, in turn, integrate the new knowledge into the evidence base to guide policy, practice and subsequent research. To this end, I encourage us to: work across disciplines more and more effectively, use grounded logic models and theories of change, engage stakeholders at all stages of the evaluation, monitor implementation of impact evaluations (make mid-course corrections when warranted), and tailor reports to specific audiences. In the words of Cynthia Osborne, "**if [you] want [your] work to be used, [you] must make it useful.**"

Broadening the focus

by Ray Pawson

Here are some suggestions on the future of evaluation inspired by my latest book *How to think like a realist*[6].

Farewell to the accreditation (or clearinghouse) model

Contrary to what was said earlier, I do not think we can go back to the accreditation model, giving verdicts on interventions. Programs, by their very nature, are **self-transformational**. Saying that a kind of program works means that it needs to be highly reproducible and to have implementation fidelity. I simply do not believe that interventions can be reproduced like that, because it is in no one's interest for interventions to remain fixed. Funders, policy makers, practitioners, and participants all have a vested interest in program adaptation. For example, the practitioner going to work in the morning is not going to say, "I want to reproduce this program", they are going to say, "I want to mess about with it, change it, apply it there"; and we cannot stop that from happening.

Another important factor is **the context dependence of effectiveness**. Effectiveness is heavily influenced by pre-existing contexts, things that happened before, or things that are in place before any intervention. And contexts are legion, anything from the cultural background, political background, economic background to the organizations that deliver interventions; all of those differ. In a way, everything works somewhere, nothing works everywhere. That is why we do what Tom Cook called contingency analysis. We ask a more complicated question: what works for whom, at what cost, in what circumstances, in what respects, with what sustainability, how implemented, and why. This is basically the way that evaluation is changing.

Farewell to the business model and atomised inquiry

We must also move from the "buisness model". The current ITT (Invitation to Tender) model, widely used for evaluation, requires key agencies to repeatedly commission external evaluations for each newly launched program. This system leads to what I refer to as atomised inquiry, where evaluations are isolated and disconnected from one another. The one-intervention-one-evaluation model has several critical flaws: it produces inconsistencies depending on the preferred evaluation paradigm, oversimplifies complex program dynamics to meet demands for clear and straightforward findings, and is constrained by funding structures that encourage premature, partial and palatable findings. Additionally, the independent nature of evaluators limits their capacity to actively shape or refine the programs they assess.

To address these limitations, the future of evaluation must shift toward building explanations of heterogenous outcomes not only within individual programs but also across families of related initiatives. By understanding patterns of success and failure across different yet comparable programs, evaluations can better inform the targeting and implementation of policies. This broader perspective enables evaluators to move beyond isolated findings and develop insights that accommodate the inherent complexity and variability of real-world interventions.

6 Pawson, Ray. How to Think Like a Realist. A Methodology for Social Science. London: Edward Elgar Publishing, 2024. https://www.e-elgar.com/shop/gbp/how-to-think-like-a-realist-9781035321094.html

Prioritise synthesis over evaluation: apply retrospective evidence prospectively

Policymakers often demonstrate a limited imagination and face resource constraints, which leads to repetitive cycles where variants of the same programs are implemented, studied, and evaluated multiple times. Despite these efforts, this repetitive process frequently fails to generate cumulative learning. Instead of learning from past evaluations, similar mistakes or oversights are repeated, creating inefficiencies in policymaking and program design.

To break this cycle, a feedback or learning loop can be established retrospectively. By synthesizing evidence from previous inquiries on comparable programs, evaluators can consolidate insights that answer the critical "conditionality question": what works, for whom, in what circumstances, in what respects, at what cost, with what sustainability, how it was implemented, and why it was effective (or not). This synthesized understanding can then serve as a foundation for developing new programs, allowing policymakers to leverage the lessons of past initiatives rather than continually starting from scratch.

Demolish the evaluation research 'silos': concentrate on commonalities

One way to achieve this synthesis is by looking at the baseline theories: what are the fundamentals of the different theories that are applied? Hence the heroic maxim that there are only three types of programs: those that supply carrots (incentives) those that supply sticks (disincentives) and those that apply sermons[7]. Incentivization, for instance, applies across all of policy domains – we incentivize people in health, in security, in independent living for disabled people – but they rely on a similar core instrument and theory. There is a lot to be gained by having a core group of evaluators that work across systems.

Begin at the beginning: Shift resources from *ex-post* to *ex-ante*

When evaluating programs, "thought experiments" often prove to be both more feasible and more valuable than real-world experiments. *Ex-ante* evaluation—evaluation conducted before program implementation—takes various forms, including scoping studies, front-end analyses, and policy scrutiny, and benefits significantly from feedback derived from research synthesis. These methods allow policymakers to anticipate potential challenges and consequences in a structured and informed manner.

However, programs are often reactive, designed in response to crises or failures. This reactionary nature frequently leads to a "do something" approach, where responsibility for implementation is handed off, and the finer details of the program are resolved only during execution. As a result, *ex-ante* considerations are often limited or neglected, leaving programs vulnerable to unforeseen barriers and unintended consequences.

A more robust model can be drawn from the legislative process. Laws, unlike programs, are durable and difficult to reverse. Before they are passed, legislation undergoes meticulous, line-by-line parliamentary debate and scrutiny. This process aims to identify and address potential loopholes, barriers, and unintended consequences prior to implementation. Applying similar rigor to program design could help ensure that initiatives are better conceived, more resilient, and more effective from the outset.

Widen the focus from "programs" to the whole policy apparatus

Program evaluation began with the attempt to evaluate interventions, specific programs in specific places. But policymaking is not just about designing interventions.

7 Matthew Hannon, Iain Cairns et al. "Carrots, sticks and sermons: Policies to unlock community energy finance in the United Kingdom", *Science* Direct, 2023.

This is in some ways a minor aspect of policymaking. There are many other public policy instruments (regulation, legislation, tribunals, management reforms, public inquiries, fiscal instruments, capacity building, etc.) that are just as interesting, potentially more powerful, and yield to the same question – "what works, for whom in what circumstances, in what respects, at what cost, with what sustainability, how implemented and why?". The research method remains the same: lay out the policy assumptions (theory) in detail and research each one.

Incorporate institutional history: learning from continuing trial and continuous error

The problem with the development of evaluation is that there are thousands and thousands of evaluations and getting to grips with them, knowing which ones to attack, which ones to synthesize is very difficult. The key policy agencies grapple with the same enduring issues, making endless tweaks and adaptations.

What happens over time, over history, matters. Mistakes are made, unintended consequences happen. And following that as a chain is an interesting way to think about evaluation. We can imagine key government agencies are tasked with improvements in specific areas. Rather than evaluating only current initiatives, historical reviews will uncover stubborn background causes, unintended consequences, unforeseen errors and partial victories. Historical analysis can build explanations by 'learning from mistakes.'

Attempt the impossible: confront the 'wicked problems'

Problems labeled as "impossible to solve" are not necessarily "impossible to understand." These so-called "wicked problems" are characterized by fundamental disagreements among key stakeholders about both the nature of the problem and potential solutions. Such issues often fall beyond the control or responsibility of any single agency and lack immediate or definitive resolutions. This complexity frequently results in policy paralysis, where no actionable steps can be agreed upon.

To address these challenges, policy thinking is increasingly adopting a "small wins" framework. Instead of relying on one-shot solutions, such as launching entirely new programs, reorganizing services, or increasing funding, this approach focuses on incremental changes. Solutions lies in shifting the narrative and subtly altering perceptions of the problem. These gradual adjustments can lead stakeholders closer to consensus and open pathways for meaningful progress.

For example, healthcare systems with universal coverage, like those in the UK, are facing relentless and unsustainable increases in demand. The overwhelmed state of mental health services highlights this strain. A potential solution lies not in expanding treatment capacity alone but in reframing the issue to promote broader societal changes. Emphasizing inclusion, rather than solely focusing on treatment, could help society better embrace individuals with neurodiversity.

Towards an evaluation culture

by Laura R. Peck

For this presentation, I consulted the source that more and more people are asking and will continue to ask: AI (or "IA" in French translation). I asked Chat GPT: What are the main issues in evaluation today? According to Chat GPT, these are the current challenges in evaluation:

- Data quality and availability
- Methodological challenges
- Causality determination
- Incorporating stakeholder input
- Interdisciplinary approaches
- Resource constraints
- Technology and innovation
- Adaptability and learning
- Ethical considerations

I will use this as a framework to help put some shape around my comments and sharing some vision for the future.

Data quality and availability

There is more and more data out there. Nevertheless, it is not all of high quality or even collected for the purpose of evaluation. This means, then, that it does not really do much to benefit evaluation or generate policy-relevant evidence for better decision-making or practice. In LIEPP's 10th anniversary roundtable[8], I made the same point about big data: that data devoid of design is close to useless. Right now, everything we do is measured and analysed, we are surveyed now more than ever before. I would put this under the label of "surveillance capitalism." For instance, I am asked to rate my "satisfaction" with nearly every customer-service interaction, almost every coffee or pharmacy purchase. I think all of you know, though, that satisfaction is not evaluative. The fact that this information is widely available does not mean that those data are any good for the purposes of better designing public policies and programs. The future of data quality and availability *for evaluation* is important. We need to be deliberate about collecting appropriate data—unbiased, comprehensive, measured, nuanced data—for evaluation purposes.

Community involvement

People who come into the evaluation field tend to do so because we want to make the world a better place. Concurrently, we tend to involve varied partners and constituents in the evaluative process as a means to that goal. Engaging people in the process of research (e.g., through community-based participatory research) and inviting people with lived expertise and learned wisdom to advise evaluation research are good practices that have been developed and are commonly deployed. That said, this approach is only a partial potential solution to embedding an equity focus in evaluation. There is much more that we can and should be doing to embed equity principles and practices into a much wider array of evaluation approaches, including those that support causal conclusions. Doing so is not yet standard practice, and we do not yet have good models for equity-transformed causal methods. I have hope that we can get there, and I have ideas for how to do so.

Data quality and availability

We are using tablets and smartphones to collect data on programs and policies, and these also have their own customized management information systems: all of this is somewhat uncoordinated and as such it does not really do too much to advance the

8 « L'évaluation, entre recherche et action », 13 May 2022, LIEPP : https://www.sciencespo.fr/fr/actualites/evaluation-entre-recherche-action/

evaluation field. This is just sort of a category of "tools." Although this is an issue for evaluation, it is not necessarily part of a solution to a vision for a better evaluation future.

Data quality and availability

Adaptability and learning relate to evaluation use. The U.S. government and some service agencies have made advances by having learning agendas and really being clear about articulating their goals for what they want to learn using evidence and funding research and evaluation. This is an important starting point for operating in an evaluation culture.

Methodological challenges and causality determination

It is my view that experimentally designed evaluations remain one of the strongest means for generating *causal* evidence about policy and program *impacts*. One major drawback that people cite is that of sample sizes being too small or samples being too idiosyncratic to be useful. This is especially relevant as we are thinking about increased interest and prioritization of impacts for gender, race, ethnicity, and intersectionally-defined subgroups. Any one evaluation is just one point of evidence. Across many evaluations, we will have many points of evidence; and there is promise to be able to produce generalizable evidence from a larger sample of those data points across more diverse samples and more contexts. Meta-evaluation and synthesis approaches have the potential to be a solution to that challenge.

Laura's evaluation dream

If imagining is *predicting*, then that was my crystal ball about where the field is going. But I am also somewhat of a dreamer. If imagining means *dreaming*, then I actually think we can improve upon *all* of the current challenges that AI suggested we face. As I think about the future of evaluation, I imagine an evaluation culture that is *not* a "feedback/satisfaction/surveillance culture," but instead one that is thoughtful about evaluation and evidence. My evaluation dream is that we can move toward an evaluation culture that is community-driven, built into program operations, matches methods to questions, leverages experimental designs for impact questions (where feasible and relevant), and uses extent and administrative data (that I hope somebody is doing something to de-bias and make more useful) or stealthfully and efficiently collected survey data ultimately to generate useful, accessible evidence to inform program improvement and policy decisions. I do feel that attaining this dream is feasible, and I am hopeful that we can overcome the feedback/satisfaction/surveillance culture and instead evolve the way that we use evidence to become an *evaluation culture*.

Le LIEPP (Laboratoire interdisciplinaire d'évaluation des politiques publiques) bénéficie du soutien du plan d'investissement France 2030 à travers l'IdEx Université Paris Cité (ANR-18-IDEX-0001).

www.sciencespo.fr/liepp

Si vous voulez recevoir les prochains échos du LIEPP et rester informés de nos activités, merci d'envoyer un courriel à : liepp@sciencespo.fr

Directrice de publication :
Anne Revillard

Edition et maquette :
Andreana Khristova
Evane Grossemy

Sciences Po - LIEPP
27 rue Saint Guillaume
75007 Paris - France
+33(0)1.45.49.83.61