Policy Evaluation and Adaptive Governance: Learning What Works, Where and Why

Michael Tan, Amara Ncube

Article type: Editorial

Volume 1, Issue 1 (2026)

Read Current Issue →

Abstract

Policy evaluation is central to accountable public management, but an average treatment effect is rarely sufficient for governing complex programmes. Decision makers need to know how an intervention produced outcomes, for whom, under which institutional conditions, at what cost, over what time period, and with what unintended consequences. This editorial argues for an evaluative model that combines causal inference with implementation analysis, theory, context, equity, cost, and organizational learning. Randomized and quasi-experimental designs remain powerful tools for estimating effects, while process evaluation, realist and theory-based approaches, qualitative inquiry, administrative-data analysis, and mixed methods can explain mechanisms and transferability. Key challenges include attribution, selection bias, implementation variation, weak outcome measures, short follow-up, missing data, publication incentives, and external validity. Priorities include linked administrative data, longer-term follow-up, replication across settings, mechanism analysis, distributional effects, cost-effectiveness, adaptive experimentation, mixed-method evaluation, and transparent reporting of null and adverse findings. Evaluation should support policy learning and adaptation rather than merely classify programmes as successful or unsuccessful.

Keywords: policy evaluation; causal inference; adaptive governance; programme evaluation; administrative data; mixed methods; evidence-informed policy

Citation

Tan, M., & Ncube, A. (2026). Policy Evaluation and Adaptive Governance: Learning What Works, Where and Why. Applied Policy, Institutions & Public Management, 1(1). https://eldranacademic.com/apipm/articles/4/

Received	Revised	Accepted	Published
27 April 2026	24 July 2026	28 August 2026	17 September 2026

Copyright

© 2026 Michael Tan, Amara Ncube. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) ↗. Use, distribution or reproduction in other forums is permitted, provided the original authors and copyright owners are credited and the original publication in this journal is cited in accordance with accepted academic practice.

Disclaimer

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, the publisher, the editors or the reviewers. Any product evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher.

1.0 Introduction

1.1 Context and Problem Framing

Evaluation is one of the principal mechanisms through which public institutions test whether programmes achieve intended results and whether scarce resources are being used responsibly. Yet the question 'does it work?' is usually underspecified. Public programmes operate across heterogeneous populations, organizations, geographies, implementation systems, and time horizons. An average effect can conceal large differences in benefit, harm, access, or cost. Experimental and quasi-experimental methods provide powerful tools for estimating causal effects when their assumptions are credible (Shadish et al., 2002; Imbens & Rubin, 2015), but causal attribution is only one dimension of policy learning.

Modern evaluation therefore draws on multiple traditions. Impact evaluation estimates changes attributable to an intervention; process evaluation examines how implementation occurred; theory-based and realist approaches investigate mechanisms and context; cost-effectiveness analysis connects outcomes with resources; and qualitative or mixed-method designs explore experience, institutional dynamics, and unintended consequences. Pawson and Tilley (1997) famously frame the realist question as what works for whom in what circumstances, while Weiss (1997) emphasizes the explanatory role of programme theory. These approaches should not be treated as competitors to rigorous causal inference. They answer different questions required for policy decisions.

1.2 Why the Issue Matters

The growing availability of administrative data and stronger causal methods has expanded what public institutions can evaluate, but methodological sophistication does not remove institutional constraints. Evaluations may occur too late, focus on easily measured outcomes, ignore implementation variation, or be designed primarily to demonstrate compliance. Deaton and Cartwright (2018) warn against treating randomized trials as automatically sufficient for causal and policy inference, while Cartwright and Hardie (2012) emphasize that transferring evidence requires understanding the support factors that allow causal mechanisms to operate. The proposition advanced here is that rigorous evaluation should combine effect estimation with mechanism, implementation, context, equity, cost, and uncertainty so that evidence can support adaptation rather than simple performance ranking.

2.0 Current Landscape

2.1 Theories of Change

Randomized evaluation remains a reference design for estimating causal effects because random assignment can balance observed and unobserved confounders in expectation. Development economics and social policy have greatly expanded the practical use of randomized field experiments, supported by methodological toolkits for design, power, treatment compliance, and outcome analysis (Duflo et al., 2007). At the same time, quasi-experimental approaches - including difference-in-differences, regression discontinuity, instrumental variables,

and matching - permit credible evaluation when randomization is infeasible or unethical. Imbens and Wooldridge (2009) synthesize major developments in this programme-evaluation literature, while Angrist and Pischke (2009) helped institutionalize design-based causal reasoning across applied social science.

Evaluation practice also increasingly recognizes implementation complexity. The Medical Research Council's guidance on complex interventions stresses the importance of process, context, and mechanisms rather than effect estimation alone (Craig et al., 2008). Moore et al. (2015) provide detailed guidance for process evaluation, including implementation, mechanisms of impact, and context. Theory-driven approaches offer complementary tools: White (2009) links theory of change to impact evaluation, Donaldson (2007) emphasizes programme theory in causal explanation, and Pawson (2013) develops realist approaches to evidence synthesis and evaluation.

2.2 Experimental and Quasi-Experimental Evaluation

Administrative data have changed the feasible scale and duration of evaluation. Government systems can provide longitudinal records on service use, employment, taxation, education, health, benefits, and enforcement, reducing reliance on bespoke short-term surveys. Connelly et al. (2016) describe the role of administrative data in the broader data revolution while also emphasizing access, linkage, quality, and governance challenges. These data are especially valuable for examining long-term outcomes and programme heterogeneity, but they require careful understanding of how administrative processes generate observations.

2.3 Monitoring Systems

Mixed methods are also receiving renewed attention because policy relevance often depends on combining causal estimates with explanations of implementation and external validity. Barnow et al. (2024) show how qualitative and mixed-method components can improve the interpretation and generalizability of impact evaluations. Greene et al. (1989) provide a foundational framework for mixed-method purposes such as triangulation, complementarity, development, initiation, and expansion. Creswell and Plano Clark (2018) provide practical design principles for integrating quantitative and qualitative evidence rather than running parallel studies that never inform one another. This logic is increasingly important when decision makers need to know not only whether an effect occurred, but how the programme operated and whether it can travel.

2.4 Organisational Learning

Institutional guidance has also moved toward evaluation systems rather than isolated studies. Gertler et al. (2016) connect impact-evaluation design to practical policy decisions about comparison groups, indicators, sampling, and implementation. Chen (2015) emphasizes theory-driven evaluation as a way to link programme design, implementation, and outcomes, while OECD (2020) documents how national evaluation systems depend on data quality, institutional mandates, capability, and use. More recently, Jacobzone et al. (2025) argue that rigorous impact evaluation and randomized trials remain underused in government despite their potential, highlighting administrative data, partnerships, and capacity as practical enablers. Evaluation quality therefore depends on both method and the institutional arrangements that commission, interpret, and act on evidence.

3.0 Critical Challenges

3.1 Attribution

Attribution remains the first methodological challenge. Non-random programme participation can produce selection bias when participants differ systematically from nonparticipants. Quasi-experimental designs can address some forms of bias, but credibility depends on design-specific assumptions and diagnostics. Heckman and

Smith (1995) show why social experiments and nonexperimental evaluations can diverge, while Imbens and Rubin (2015) emphasize explicit potential-outcomes reasoning. Evaluation reports should therefore make identification assumptions visible rather than presenting statistical adjustment as automatic causal proof.

Implementation variation creates a second challenge. A programme assigned uniformly may be delivered differently across sites, staff, or population groups. Average effects can then combine heterogeneous interventions. Moore et al. (2015) show why fidelity, dose, reach, mechanisms, and contextual factors should be measured alongside outcomes. If implementation is weak, a null effect does not necessarily invalidate the policy theory; if implementation is highly adapted, a positive effect may not be attributable to the nominal programme as originally specified.

3.2 Contextual Dependence

External validity is a third challenge. Results from one jurisdiction, population, or implementation system may not transfer to another. Leviton (2017) argues that generalization requires explicit attention to context and programme theory, while Vivalt (2020) demonstrates empirically that effect sizes can vary substantially across settings. Stuart et al. (2011) and Tipton (2013) provide statistical approaches for assessing or improving generalizability from experimental samples. These tools are valuable, but transfer ultimately requires substantive knowledge of mechanisms and institutional conditions.

Outcome selection and follow-up periods create a fourth challenge. Evaluations often privilege outcomes that are administratively available or responsive within funding cycles, even when policy goals are long term. Short follow-up can miss delayed effects, adaptation, displacement, or harm. Administrative data can extend follow-up, but measurement validity remains crucial because administrative indicators may reflect eligibility rules, reporting practices, or enforcement intensity rather than underlying outcomes. Connelly et al. (2016) therefore emphasize that large administrative datasets do not eliminate measurement problems.

3.3 Implementation Fidelity

Political and publication incentives create additional distortions. Decision makers may prefer evaluations that confirm programme success, while researchers and journals often have stronger incentives to publish statistically significant findings than null or adverse results. Sanderson (2002) argues that evaluation should contribute to policy learning rather than function only as technical performance measurement. Adaptive governance requires evidence systems in which disappointing results can trigger redesign without automatically becoming institutional failure. This is especially important for pilot programmes where learning is the purpose of experimentation.

3.4 Balancing Adaptation with Comparability

Cost and distribution are frequently underdeveloped dimensions. A programme may produce an average benefit but be prohibitively expensive, shift costs to another agency, or improve outcomes for already advantaged groups while leaving others behind. Policy evaluation should therefore examine resource use, opportunity cost, and distributional effects alongside impact. Decision-focused evaluation frameworks increasingly argue that causal effects, cost, mechanisms, implementation, and uncertainty must be considered together rather than sequentially.

4.0 Discussion and Priorities

4.1 Integrative Discussion

Rigorous policy learning requires an evaluative portfolio rather than a hierarchy in which one design answers every question. Randomized and quasi-experimental methods are strongest when the central uncertainty concerns attribution. Process and qualitative methods are stronger when the uncertainty concerns implementation, experience, or mechanism. Realist and theory-based approaches clarify causal assumptions and contextual dependencies. Cost and decision analysis connect effects to resource constraints. Mixed methods can integrate these strands when a decision requires several types of evidence.

This integration also changes how success should be understood. A programme with a modest average effect may be highly valuable for a priority subgroup; a programme with a strong effect may be unsustainable at scale; a null result may conceal successful implementation in some settings and failure in others. Deaton and Cartwright (2018) caution that experimental estimates do not automatically identify the causal structure needed for external policy decisions. Policy learning should therefore distinguish effect estimation from explanation and transfer.

Adaptive governance provides the institutional rationale for this broader evaluation model. Evaluation should create feedback that can alter programme design, targeting, delivery, and resource allocation. Patton's (2011) developmental evaluation emphasizes learning in complex and changing environments, while Sanderson (2002) connects evaluation to learning under uncertainty. Manski (2013) adds an important decision-theoretic perspective: public choices frequently must be made under irreducible uncertainty rather than postponed until evidence becomes definitive. The goal is not perpetual experimentation without accountability, but structured adaptation in which changes are evidence informed, documented, re-evaluated, and explicit about the uncertainty decision makers still face.

Complementary scholarship informing this synthesis includes OECD. (2020); and Patton (2011). These works broaden the evidentiary base considered across the field.

4.2 Research and Practice Priorities

First, linked administrative data should be used more systematically for long-term and population-level evaluation. Researchers should document data-generating processes, linkage quality, missingness, coding changes, and legal governance rather than treating administrative records as neutral observations. Longitudinal linkage can reveal persistence, delayed effects, service substitution, and downstream costs that short studies miss.

Second, replication across settings should become a normal part of policy evaluation. Replication can distinguish site-specific success from robust programme mechanisms and can support more credible external validity. Researchers should report differences in institutions, populations, workforce, implementation, and economic conditions rather than describing replication simply as repetition.

Third, mechanism and implementation measurement should accompany effect estimation. Programme theory should specify the intermediate changes expected to produce outcomes, and process evaluation should test whether those changes occurred. White (2009) and Moore et al. (2015) provide complementary approaches for connecting causal pathways with implementation evidence.

Fourth, distributional analysis should be planned prospectively. Average treatment effects should be supplemented with theoretically justified subgroup and equity analyses where sample size permits. Researchers should avoid indiscriminate subgroup mining while still examining whether benefits and burdens differ by geography, income, age, disability, or other policy-relevant characteristics.

Fifth, cost-effectiveness and opportunity costs should be integrated earlier. Programme costs should include administration, staffing, training, infrastructure, monitoring, and participant burdens where relevant. Comparative evaluations should use realistic alternatives, including existing practice, rather than comparing new interventions only with no action.

Sixth, adaptive experimentation and mixed-method evaluation should be used where policies evolve during implementation. Sequential trials, phased rollouts, quasi-experimental changes, and embedded qualitative inquiry can create learning while programmes operate. Barnow et al. (2024) show how mixed methods can increase policy relevance, particularly for external validity.

Finally, transparent reporting of null, heterogeneous, and adverse findings should be treated as a public good. Preregistration, analysis plans, complete outcome reporting, and publication of credible null results can reduce selective learning. Evaluation systems should reward evidence that improves decisions, even when it contradicts the programme's original expectations.

4.3 Implications for the Journal's Scope

Policy-evaluation scholarship should contribute to institutional learning rather than merely label programmes as successes or failures. Valuable contributions include randomized and quasi-experimental studies, natural experiments, administrative-data analyses, implementation and process evaluations, realist and theory-based work, cost-effectiveness studies, mixed methods, and research on evaluation use within public organizations.

The strongest studies make assumptions explicit and align methods with decisions. They explain what outcome was measured, what counterfactual supports the causal claim, how implementation varied, whose outcomes changed, what the intervention cost, and what limits transfer to other settings. Null and adverse findings deserve publication when they are methodologically credible and analytically informative because policy learning depends as much on identifying limits as on confirming benefits.

4.4 Conclusion

Policy evaluation should answer more than whether an average effect can be detected. Public managers need evidence about mechanism, implementation, context, equity, cost, time, and unintended consequences if they are to decide whether programmes should be expanded, modified, targeted differently, or discontinued.

No single method can supply all of that information. Causal designs, process evaluation, theory, qualitative research, administrative data, cost analysis, and mixed methods should be combined according to the decision problem. Evaluation becomes most valuable when it supports disciplined adaptation: learning what works, where, for whom, why, and at what cost, while remaining transparent about uncertainty and failure.

5.0 Declarations

5.1 Author Contributions

The authors contributed jointly to the conceptual development, literature synthesis, critical interpretation, drafting, and revision of the editorial.

5.2 Funding

No external funding was received for the preparation of this editorial.

5.3 Competing Interests

The authors declare that they have no conflicts of interest or competing interests relevant to this editorial.

5.4 Data Availability

No original datasets were generated or analysed for this editorial.

5.5 AI Use Disclosure

Generative AI was used to support literature organization and language drafting during preparation of this editorial. The authors remain responsible for independently verifying the accuracy of the arguments, citations, references, and final submitted text in accordance with the publisher's policy.

6.0 References

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Barnow, B. S., Pandey, S. K., & Luo, Q. E. (2024). How mixed-methods research can improve the policy relevance of impact evaluations. Evaluation Review, 48(3), 495-514. https://doi.org/10.1177/0193841X241227480

Cartwright, N., & Hardie, J. (2012). Evidence-based policy: A practical guide to doing it better. Oxford University Press.

Chen, H.-T. (2015). Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective (2nd ed.). Sage.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social Science Research, 59, 1-12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. BMJ, 337, a1655. https://doi.org/10.1136/bmj.a1655

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Sage.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2-21. https://doi.org/10.1016/j.socscimed.2017.12.005

Donaldson, S. I. (2007). Program theory-driven evaluation science: Strategies and applications. Lawrence Erlbaum Associates.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz & J. A. Strauss (Eds.), Handbook of development economics (Vol. 4, pp. 3895-3962). Elsevier. https://doi.org/10.1016/S1573-4471(07)04061-2

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). Impact evaluation in practice (2nd ed.). World Bank.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. Educational Evaluation and Policy Analysis, 11(3), 255-274. https://doi.org/10.3102/01623737011003255

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. Journal of Economic Perspectives, 9(2), 85-110. https://doi.org/10.1257/jep.9.2.85

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47(1), 5-86. https://doi.org/10.1257/jel.47.1.5

Jacobzone, S., Picalarga, S., Dynes, L., Williams, E., Greenwell, H., & Bowers, P. (2025). Unleashing the policy potential of rigorous impact evaluation and randomised trials. OECD Working Papers on Public Governance. https://doi.org/10.1787/5c135873-en

Leviton, L. C. (2017). Generalizing about public health interventions: A mixed-methods approach to external validity. Annual Review of Public Health, 38, 371-391. https://doi.org/10.1146/annurev-publhealth-031816-044509

Manski, C. F. (2013). Public policy in an uncertain world: Analysis and decisions. Harvard University Press.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. BMJ, 350, h1258. https://doi.org/10.1136/bmj.h1258

OECD. (2020). Improving governance with policy evaluation: Lessons from country experiences. OECD Publishing. https://doi.org/10.1787/89b1577d-en

Patton, M. Q. (2011). Developmental evaluation: Applying complexity concepts to enhance innovation and use. Guilford Press.

Pawson, R. (2013). The science of evaluation: A realist manifesto. Sage.

Pawson, R., & Tilley, N. (1997). Realistic evaluation. Sage.

Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. Public Administration, 80(1), 1-22. https://doi.org/10.1111/1467-9299.00292

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A, 174(2), 369-386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics, 38(3), 239-266. https://doi.org/10.3102/1076998612441947

Vivalt, E. (2020). How much can we generalize from impact evaluations? Journal of the European Economic Association, 18(6), 3045-3089. https://doi.org/10.1093/jeea/jvaa019

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. New Directions for Evaluation, 1997(76), 41-55. https://doi.org/10.1002/ev.1086

White, H. (2009). Theory-based impact evaluation: Principles and practice. Journal of Development Effectiveness, 1(3), 271-284. https://doi.org/10.1080/19439340903114628