



Experiments in government performance

by Rachel Gisselquist and Miguel Niño-Zarazúa

FINDINGS

RCTs have provided solid evidence for the effectiveness of certain kinds of interventions, including conditional cash transfers, information campaigns, specific financial incentives, and public deliberation activities

They are, however, unable to shed light on many other types of development issues due to their limited capacity to randomize policy decisions

The tight time frames of RCTs also mean that impact cannot always be captured

RCTs tend to have strong internal validity, documenting the effects of changes under set conditions well, but suffer from weak external validity, posing challenges for the generalizability of findings

Despite their benefits, RCTs carry certain limitations and, as a result, need to be used in conjunction with other methodologies for sound policy analysis

Experimental studies using randomized controlled trials (RCTs) are a powerful tool in policy analysis. They have been sometimes hailed as the best means of identifying ‘what works’ in development policy. However, it would be unwise to rely solely on findings from RCTs to guide policy. While RCTs work well to evaluate particular types of interventions and development theories, they tell us little about the alternatives – which are numerous, particularly in certain areas, such as governance reform.

Theories of government performance

Analysis of how and how well governments perform is central to the study of politics, and the literature suggests a vast array of structural, institutional, cultural, and individual factors. Institutional theories, for instance, point to specific reforms of the electoral system and the decentralization of administrative responsibilities as means to broaden popular representation, increase public accountability, and improve public service delivery. Theoretical work also highlights the decisive role of leaders in government performance, suggesting the value of international policies in incentivizing ‘better’ choices by national leaders and support for institutions such as the media that keep leaders in check.

We conducted a systematic review of experimental and non-experimental work on government performance, which shows that experimental approaches have played an important role in addressing some of these issues. Considering improved government performance in terms of better provision and use of public services, related welfare outcomes, and the performance of public sector employees, analysis shows support across multiple experimental studies for the following types of interventions:

- antipoverty policies such as conditional cash transfer programmes
- campaigns providing information about government services and citizen rights
- financial incentives and relatively minor administrative reforms that change incentives for civil servants
- public deliberation activities at the local level and community-based monitoring initiatives

These findings are worthy of careful attention, but it is equally important to recognize the limitations of experimental approaches, three of which are outlined below.

RCTs do not – and cannot – address many important questions in development

Interventions shown to be effective by RCTs tend to be of a particular type: they involve relatively small-scale efforts – such as the provision of information



Receiving cash transfer payments. © Dominic Chavez/World Bank

and/or material incentives – to influence decision-making by relatively ‘average’ individuals. Since experiments must be conducted on large numbers of equivalent units in order to gain precise estimates, they work best in analyzing factors that can be randomized at the individual or household level, or at another relatively low level of aggregation such as the village. Experiments also are limited in the sorts of factors that they can manipulate for practical and ethical reasons; for instance, it would not generally be possible to play around with electoral rules, randomizing them across constituencies.

Experiments can sometimes push the boundaries of these limitations, for instance through innovative collaborations with governments. One example is the Progresar/Oportunidades programme in Mexico, as its staggered rollout made it possible to use an experimental design to assess its impact. The identification of ‘natural’ experiments – which tend to be rare – can also make possible analysis of higher-level factors using experimental methods.

later – while missing its impact on political stabilization in subsequent decades.

To address this limitation, experiments could be run for longer periods and include multiple intermediate periods of analysis. Costs, however, can be prohibitive. One innovative way to minimize costs is to do ‘ancillary experiments’, drawing on existing experimental data to investigate new questions. The challenge, however, remains that wider time windows allow wider ranges of developments that potentially contribute to measured changes, thus weakening causal inference.

They have weak insight into how findings can be generalized

To support evidence-based policy making, RCTs should provide rigorous evidence that an intervention has had a given impact and that it can be expected to have a similar impact in other contexts. RCTs tend to do well in the first area but poorly in the second – i.e., they have strong internal validity but weak external validity. Experimental work has made strides towards strengthening the external validity of its findings, in particular by replicating interventions in multiple contexts, and conducting systematic reviews and meta-regression analyses. Experimental studies could further strengthen external validity by speaking more directly to broader theoretical propositions and by explicitly considering findings in terms of the contextual factors that may influence them.



© Salahaldeen Nadir / World Bank

They do not tell us about impact beyond a relatively narrow time window

Most RCTs hinge on a comparison of measures between treatment and control groups before an intervention takes place and some months – or occasionally a year or two afterwards. Yet, in terms of government performance, it is not clear that ‘impact’ becomes evident within this narrow time window. Many theories of government suggest that change occurs over years, decades, and even generations.

Moreover, impact is not necessarily linear; measuring the trajectory between two points in time can be misleading. For example, if the relationship between economic liberalization and political stability is in fact ‘J-shaped’, an RCT could blame economic liberalization for causing political instability a year

Methodological eclecticism

RCTs are a powerful methodological tool, which can be further honed. But, they are not the best tool for all purposes. Other rigorous quantitative approaches that should be included as part of the portfolio of impact evaluation methods include, among others, instrumental variables, regression discontinuity, direct matching, propensity score matching, linear regression, and difference-in-differences. Qualitative methods are also relevant for addressing specific questions and these may include comparative case studies, participant observation, interviews, focus groups, and historical process tracing. Each of these tools has comparative strengths and weaknesses for policy analysis.

The real gold standard for evidence on government performance, thus, is not RCTs, but methodological eclecticism.