World Yearbook of Education 2019

Comparative Methodology in the Era of Big Data and Global Networks

Edited by Radhika Gorur, Sam Sellar and Gita Steiner-Khamsi

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Part 1

Impacts

Randomized Controlled Trials: league leader in the hierarchy of evidence?

Gita Steiner-Khamsi

Randomized Controlled Trials (RCTs) are revered by some as the "gold standard" for assessing the impact of interventions or reforms and condemned by others for reducing the complexity of cause-effect relations to a few controllable and measurable conditions. The method has experienced a popularity to the extent that policymakers nowadays do not dare to allocate funds and make political decisions without the scientific stamp of approval provided by a randomized experiment. Spearheaded in its infancy by renowned development economists, such as Jeffrey Sachs, Earth Institute, Columbia University, and Esther Duflo, co-founder and director of the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology, the method is nowadays considered a *dictum* for externally funded programs and reform projects of developing countries. In fact, within a short period of time, many funding agencies and donors have come to consider RCTs as the only reliable method for assessing whether their financial investment was worth the actual outcomes. As a corollary, the two contributions selected for this part of the book draw on case studies of RCTs in rural China as well as a comparative RCT study in Kenya and Uganda to critically reflect on the method.

RCTs are nowadays produced on grand scale, in particular, on topics that are controversial, such as studies that measure the impact of contract teachers or the impact of low-fee private schools on student achievement (Klees, 2018; Romero, Sandefur, & Sandholtz, 2017). The choice of method is not coincidental. RCTs have achieved such an authoritative status to afford policymakers with the necessary justification to make difficult decisions on controversial issues.

For this book, two internationally renowned scholars, Prashant Loyalka (Stanford University) and Moses Oketch (University of London), discuss the opportunities and shortcomings of RCTs, randomized experiments, or impact evaluations. Both write from extensive experience. What they present are sophisticated accounts of why policymakers are enamored with impact evaluations and how policy analysts design an RCT to draw policy-relevant conclusions. Both Chapters 2 and 3 discuss the opportunities and shortcomings of RCTs.

One of the criticisms, succinctly summarized and illustrated by Loyalka in Chapter 2 is "that RCTs are atheoretical, only explaining 'what works' and not why" (Loyalka, Chapter 2). In an attempt to overcome this shortcoming, the

research team focused on causal chain effects, which they managed to identify based on a thorough review of theory debates on the topic. Loyalka specifies the RCT design for the Rural Education Action Program (REAP) in China to illustrate how RCTs may be nested in theory, and in fact contribute and advance theory debates, *if* adequately designed. In REAP, the method was used with the intention to gain insights on how to improve policies surrounding human development (education, health and nutrition) in China. The program was not only large and comprehensive, spanning over several sectors, but the timing of the impact evaluation was meaningful. Essentially, REAP introduced and disseminated RCTs in the education sector of China. By the year 2018, over 80% of published impact evaluation studies on education in China were published by REAP scholars, many of whom were supervised by Prashant Loyalka. His deliberations in Chapter 3 focus on a randomized experiment that sought to evaluate the impact of the teacher professional development program on student achievement. Approximately 600 teachers and their approximately 33,000 students were randomized to different treatment and control conditions. The finding was unexpected: the RCT indicated that teacher's professional development did not have an impact on student achievement. In most cases, the mandate of an RCT researcher ends here: to provide robust evidence to planners and funders for a "stop" or "go" verdict. This was not the case with this particular REAP RCT. Rather than ending the investigation with the proof of "no impact" and moving on to orchestrating the next RCT study, the unexpected finding triggered a host of follow-up questions that led the research team to investigate causal chain effects. In particular, why and under what conditions is a teacher professional development program effective? The scope of the follow-up research questions and hypotheses were guided by a systematic literature review on the topic. For example, the research literature suggested that follow-up mentoring and testing of teachers matter a great deal or that the quality of a professional development program determines what teacher education students learn and translate in their own pedagogical practice. In other words, the unexpected finding of the RCT forced the research team to dig deeper and investigate the features of the training programs in order to explain the varied impact of teachers' professional development on students. In a next step, the research team tested the key features of professional development programs. Loyalka's research team also considered secondary effects of professional development of teachers and acknowledged that teacher professional development perhaps did not have a direct impact on student learning outcomes but rather had other positive effects on students. In other words, different from what RCTs are notorious for – tapping in the dark or in a theory-free zone – the design of this particular RCT became theory driven. The method was used to disentangle the various features of a professional development program, to test theory and to understand the complexity and fuzziness of causal chain effects.

Whereas Loyalka's chapter is an invitation to move beyond superficially using RCT for "stop" or "go" policy decisions, in Chapter 3 Oketch criticizes the meaninglessness of RCTs, that is, the complete disregard for context, system or macro-level factors.

Oketch convincingly argues that by design the narrow focus on program-level variables generates a blind spot for system-level analyses. After all, one and the same program or intervention means something different, and is translated and recontextualized differently, in two different educational systems. In other words, "external conditions" (Oketch, Chapter 3) matter. To make the case, Oketch presents the comparative RCT study of East Africa Quality in Early Learning (EAQEL) that he and his associates carried out in Kenya and Uganda. Similar to the REAP, the EAQEL program intended to improve student learning. There were three program components that the RCT tested in terms of their impact on learning outcomes: (i) teacher preparedness and practice, (ii) school leadership and (iii) classroom learning environments. The study randomly divided the samples into a control group (no exposure to the EAQEL program), a Core Model and a Core Model Plus. Oketch summarizes the different features of the two EAQEL Core Models as follows:

> The Core Model involved early grade teachers being trained on the instructional approach, which was child centered, systematic and focused on social interaction. In addition, schools were supported to improve teachers' and pupils' access to and use of appropriate teaching and learning materials. Project technical staff worked with head teachers, key teachers and district education staff from decentralized teacher support resource institutions to train teachers and provide in-class mentoring support. The Core Model Plus included all of the aspects of the Core Model and a parental involvement component. The aim was to encourage literacy by establishing mini-libraries in selected homes, and encouraging parents to borrow books, read to their children and tell stories to their children.
>
> (Oketch, Chapter 3)

As with all RCTs, great attention was paid to selecting comparable samples in the two countries (districts with low-performing schools) and to conducting a baseline study on student performance *before* the EAQEL program was launched. As expected, the baseline study confirmed the low learning levels in the selected districts and schools of both countries. Also as expected, the additional parental involvement component yielded positive results. What was not accounted for, however, were the significant differences in student learning outcomes in the two countries or educational systems. Even though the students were exposed to the identical EAQEL program, the broader policy context of Uganda was apparently more amenable to a program that targeted the improvement of numeracy, oral and written literacy for grade 1, 2 and 3 students. Students in Uganda outperformed students in Kenya in all domains. However, a more nuanced picture emerged when the research team differentiated between schools in terms of how well they have implemented the program. The students in Uganda still outperformed their peers in Kenya, but a least the EAQEL had a significant positive impact on learning outcomes in high-implementing schools of Kenya. Oketch draws on the EAQEL RCT example to highlight the importance of policy context, system performance,

or external conditions that affect what is going on in a school. RCTs by design are too micro- and program-specific to enable a meaningful interpretation of why students in one district or one country benefit more from a program (such as the EAQEL) than students in another.

As an alternative to program-level RCTs, Oketch proposes to think bigger and apply a holistic system-oriented approach that enables the researcher to dig deeper into the interdependence between various mechanisms of change or reform. An investigation of why reforms have not been taken or were not fully implemented in one context, as opposed to another context, would contribute to a better understanding of "reform pathways" and helps to explain why one and the same reform, regardless of the individual features of the reform program, has a different impact in different educational systems.

Regardless of whether RCT design should be more theory based, allowing for more depth, as proposed by Loyalka (Chapter 2), or more contextual, taking into account system differences (Chapter 3), the method has raised enormous expectations among policymakers and analysts that are nearly impossible to fulfill. Arguably, the unrealistic expectations are self-induced. Banerjee and Duflo's claim is that RCTs provide "gold standard" evidence for policymakers (Banerjee & Duflo, 2011). The superiority of RCTs over any other method of inquiry for making causal claims and assessing the effectiveness of a program has also been stated by the prominent econometrician Guido W. Imbens (2010, p. 407):

> [R]andomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top.
>
> (cited in Oketch, Chapter 3)

This part of the book may be read as proposition to critically reflect on the hierarchy of evidence on one hand, and on the politics and economics of RCTs on the other. What counts as evidence, whose knowledge counts as evidence, and how is the discursive power or authoritative status of evidence established and reproduced? Furthermore, the high authoritative status of RCTs begs the question of what political and economic gains policymakers possibly associate with RCTs. There exists a fascinating body of literature on "evidence" and "best practices" that is worth mentioning here.

We have engaged in a critical examination of evidence, typically seen as the foundation for knowledge-based policy regulation, in another study (see Baek et al., 2018). We found it essential to examine in greater depth the question of what counts as evidence and how it functions. A few findings of the literature may be worth reiterating here.

Previous literature has acknowledged the multiple types of evidence used in the policy process (Davies, Nutley, & Smith, 2000; Weiss, 1979; Kvernbekk, 2011). Evidence ranges from research findings, existing statistics, to expert knowledge and secondary sources. Despite the broad definition of evidence, what constitutes "good" evidence has been extensively debated. Our review of the literature suggests that the definition of evidence (i) changes over time, (ii) is context specific

and (iii) varies depending on the stage in the policy process. First, Hadorn and his colleagues (1996) point to the hierarchies of evidence whereby some forms are perceived as more robust than other forms. For example, in some countries RCTs now seem to rank higher than expert opinions. If observed over a period of several decades, one would most likely find that some types of evidence come into fashion, whereas others burn out over time. Second, Hulme, Hulme, and Rauschenberger (2017) examine how the global script of evidence-based educational reform is locally adapted, recontextualized, or selectively borrowed in three different policy environments of Great Britain. In all three cases, there is a commitment to learning from "what works," but its translation into the policy contexts of Scotland, England and Wales differs greatly. For example, the choice of RCTs to identify "best practices" is only found in the What Works Centres in England. Policy analysts in Scotland and Wales tend to use other tools for determining "what works." Finally, McDonnell and Weatherford (2013) expand this argument and claim that policy actors use different types of evidence for each stage of policy process. During the problem-definition and solution-identification stage, non-research evidence such as anecdotes and metaphors are used to humanize the problem by appealing to policy actors' and the public's core values, in addition to research-based evidence. In the policy design stage, evidence is more technical and less emotional and normative. Finally, I would like to add here that at the policy evaluation stage, empirical research is used to justify whether a reform should be continued, discontinued or modified.

This is where the politics and economics of RCTs come into play. It appears that policymakers tend to choose RCTs as their preferred method of empirical inquiry (i) if a program or reform is contested and they need to rely on RCTs' authoritative status as the "gold standard" for making "stop" or "go" decisions, and/or (ii) if they plan to scale up a program or reform nationwide, or in the case of international organizations, globally. The latter explains why enormous human and financial resources are channeled into the expensive and labor-intensive RCTs. The return on investment is made in form of transferring the same program to other schools, districts, regions or countries.

The two chapters in this part demonstrate masterfully the methodological limitations of RCTs for "stop" or "go" decisions (Loyalka, Chapter 2) and for scaling up program, reforms or "interventions" regardless of " 'external conditions', namely context, system and reform capacity" (Oketch, Chapter 3).

References

Baek, C., Hörmann, B., Karseth, B., Pizmony-Levy, O., Sivesind, K., & Steiner-Khamsi, G. (2018). Policy learning in Norwegian school reform: A social network analysis of the 2020 incremental reform. *Nordic Journal of Studies in Educational Policy*, *4*(1), 24–37. doi:10.1080/20020317.2017.1412747

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, NY: Public Affairs.

Davies, H., Nutley, S., & Smith, P. (Eds.). (2000). *What works? Evidence-based policy and practice in public services*. Bristol: Policy Press.