Received: 12 March 2019 | Revised: 4 October 2019 | Accepted: 7 November 2019

DOI: 10.1111/1468-4446.12723

SPECIAL ISSUE

WILEY

What's on trial? The making of field experiments in international development

Luciana de Souza Leão

Department of Sociology, College of Literature, Science and the Arts, University of Michigan, Ann Arbor, MI

Correspondence
Luciana de Souza Leão, Department of Sociology, College of Literature, Science and the Arts, University of Michigan, 500 South State Street, Room 3224, Ann Arbor 48109, MI, USA.
Email: lsleao@umich.edu

Abstract

In the last 20 years, the drive for evidence-based policymaking has been coupled with a concurrent push for the use of randomized controlled trials (RCTs) as the "gold-standard" for generating rigorous evidence on whether or not development interventions work. Drawing on content analysis of 63 development RCTs and 4 years of participant observation, I provide a rich description of the diverse set of actors and the transnational organizational effort required to implement development RCTs and maintain their "scientific status." Particularly, I investigate the boundary work that proponents of RCTs—also known as *randomistas*—do to differentiate the purposes and merits of testing development projects from doing them, as a way to bypass the political and ethical problems presented by adopting the experimental method with foreign aid beneficiaries in poor countries. Although *randomistas* have been mostly successful in differentiating RCTs from the projects evaluated, I also examine cases where they were not able to do so, as a means to highlight the controversies associated with implementing RCTs in international development.

KEYWORDS

economics, field experiments, global poverty, international development, NGOs

1 | INTRODUCTION

In the last 20 years, the drive for evidence-based policymaking has been coupled with a concurrent push for the use of randomized controlled trials (RCTs) as the "gold-standard" for generating rigorous evidence for whether or not development interventions work. Initially restricted to a handful of researchers located in Poverty Labs within economics departments in the US, RCTs are now being used to test almost everything from strategies to reduce electoral corruption in Sierra Leone to microcredit programs in Peru (Banerjee & Duflo, 2011). In the development community, if researchers, governmental officials, and donors want to know "what really works", it is widely accepted that RCTs must be implemented to avoid a naïve or biased answer (Deaton, 2010; Harrison, 2011). Furthermore, the growing institutionalization of the method is evidenced by wide coverage in the press and the conferment of several of the most prestigious awards for contributions to economics on Esther Duflo, a leading development economist and RCT advocate (Ogden, 2016).

Development RCTs institute experiments in everyday life to measure the impact of different poverty-alleviation policies by comparing the results of treatment and control groups. A 1997 primary school deworming RCT is representative of this type of experimentation (Miguel & Kremer, 2004). In this RCT, the goal was to estimate the effect of intestinal diseases on educational outcomes for young children. Researchers selected 50 schools in rural Kenya to receive deworming medicine for free, and 25 schools were selected as controls that initially did not participate in the program. Researchers compared educational outcomes in the control and treatment groups and found considerable improvements in test scores and school attendance not only for students in treatment schools, but also spillover effects for kids that did not receive treatment. This finding led to a policy recommendation of distributing school-based deworming pills throughout the developing world, and similar programs had reached over 285 million children by 2017 (JPAL, 2017). Experiments like this are now implemented in distinct policy areas trying to answer a variety of development questions. In common, they use the comparison between control and treatment groups to test the effectiveness of development projects in countries from the Global South.

How are US-based researchers able to implement field experiments in developing countries and to persuade others about their "scientific status"? Similar to other forms of field experimentation, the implementation of development RCTs requires ongoing negotiations between the need to control the "messiness" of the field for scientific purposes, while guaranteeing the cooperation, access, and buy-in of local populations so that the experiment can happen (Henke, 2000; Kohler & Vetter, 2016). As field-sites, however, developing countries present particular challenges for the successful balance between control and cooperation. These are places where any type of foreign intervention is inevitably entangled in the controversial politics of the foreign aid industry (Escobar, 2012; Ferguson, 1990), making the distinction between development practice and research particularly blurry (Rayzberg, 2019a). How do development economists bypass the ethical and political controversies associated with foreign aid and convince others that developing countries can serve as appropriate sites to test economic theories?

To answer these questions, I adopt concepts from science and technology studies to describe the network that needs to be in place to implement development RCTs and to make them scientifically and politically plausible, reproducible, and disseminated (Eyal, 2013). Particularly, I highlight the multiple actors involved in implementing these experiments and the complex chains of transcriptions required to generate data from the messy reality of the field and turn it into something useful for both academics and policy officials (Latour, 1999). I demonstrate that proponents of development RCTs, or *randomistas* as they became known (Deaton, 2010), build on the ambiguity of what is being tested—is it an economic theory, a development project, or both?—to navigate the political, practical, and ethical problems associated with development aid and with randomly assigning social policy beneficiaries to treatment and controls groups. In doing so, I also argue that the scientific success of development RCTs is contingent on their ability to construct the image of "the field" as a place that is free of politics and bureaucratic interference. For the most part, *randomistas* have been successful in creating a boundary between the purposes and merits of testing development projects and doing them. However, I also examine cases where they were not successful in order to highlight the controversies associated with implementing RCTs in international development.

In recent years, a number of social scientists have started to problematize development RCTs, mostly by highlighting the strategies that *randomistas* adopt to transform contested development questions into seemingly technical problems (e.g., Berndt, 2015; Deaton & Cartwright, 2016; Rayzberg, 2019a). This paper contributes to this scholarly conversation about development RCTs in two key ways. First, by combining content analysis of 63 RCTs and ethnographic data collected during 4 years of fieldwork, it provides a systematic empirical account of how these experiments operate on the ground. This account highlights the diversity of actors involved in RCTs, while so far attention has been given exclusively to the *randomistas* themselves (exceptions are Kabeer, 2019; Rayzberg, 2019b). Second, this paper unveils the processes through which development RCTs can lose their scientific status, helping to elucidate the problematic nature of RCTs from the viewpoint of the actors most affected by these experiments. In doing so, it contributes to a broader understanding of the politics of testing in international development and the inequalities inherent to the diffusion of global evaluation standards.

The article is organized as follows. The first section explains what development RCTs are and illustrates how they differ from previous experiments done in economics. Second, I present my methods and data. Relying on ethnographic data, in the third section, I describe the organizational network that needs to be in place to make development RCTs work and how actors involved in this network deal with controversies associated with RCTs to make them credible to academic and policy audiences. Fourth, I explain the strategies that *randomistas* adopt to bypass the problematic nature of using the experimental method in developing countries, and I address the main ways that development RCTs are contested. Finally, I conclude by drawing implications from my case for a critical evaluation of testing.

2 | PREVIOUS EXPERIMENTATION IN ECONOMICS AND THE NOVELTIES OF DEVELOPMENT RCTS

Experimental trials in economics have a long tradition. In the 1960s and 1970s, the Negative Income Tax Trials used the experimental method to test the effectiveness of different social policies and had long-term effects on welfare debates in the United States (Rogers-Dillon, 2004). Likewise, during the 1980s and 1990s, economic research took the form of laboratory experiments, testing different hypotheses on behavioral economics that were highly influential for the conceptualization of electronic markets and behavioral theory (Guala, 2007). The development RCTs and Poverty Labs that are the object of this study, even if they build on the prestige of these previous experiments, differ from early economics experiments in two ways.

First, in contrast to most behavioral economics research that takes the form of laboratory experiments (in which volunteers enter a research lab to make decisions in a controlled environment), development RCTs function as field experiments: they are not only performed in the field, but also the division between control and treatment groups is done with real people, schools, and communities living their everyday lives. As *randomistas* themselves point out: “There may be more to learn about human behavior from the choices made by Kenyan farmers confronted with a real choice than from those made by American undergraduates in laboratory conditions” (Duflo, 2003, p. 8). This means that development RCTs, at least in their design, do not have to deal with criticism about “mock settings” or “stage action”: the experiments take place in situ (MacKenzie, Muniesa, & Siu, 2007). Instead, similar to other field sciences, they face a different type of credibility challenge, namely, the need to continuously demonstrate that experiments in developing countries could retain certain characteristics of lab sciences, such as generating generalizable, “placeless knowledge and being inconsequential” (Guggenheim, 2012, p. 102; Kohler & Vetter, 2016).

Second, while development RCTs and the social experiments that took place in the United States in the 1960s similarly happen in the field, they differ in scale, objectives, and geographical reach. In a recent publication from the main Poverty Lab (JPAL), researchers differentiate their experiments from the past ones on the following basis:

> *Unlike the early “social experiments” conducted in the United States…many of the RCTs that have been conducted in recent years in developing countries have had fairly small budgets, making them affordable*

> *for development economists. Working with local partners on a smaller scale has also given more flexibility to researchers, who can often influence program design. As a result, RCTs have become a powerful research tool.* (Duflo et al., 2007, p. 3)

"Researchers who can often influence program design." This comment points to the key difference between development RCTs in the 2000s and earlier "social experiments" in the United States: while the latter partnered with US government agencies, current RCTs are implemented together with non-governmental organizations (NGOs) and foreign aid agencies in developing countries, allowing for much more flexibility to use the experimental method in field conditions than before (de Souza Leao & Eyal, 2019). Hence, while the 1960s' social experiments could not assign participants randomly to a "no-treatment" control group and used the non-sampled population as their implicit control for comparisons (Riecken & Boruch, 1975), development RCTs seek to randomly assign beneficiaries to control groups, attempting to portray in this way an image of greater scientificity than previous experiments. Yet, since development RCTs are implemented in remote areas of developing countries, this also means that these experiments lack administrative data that is available in developed countries; they have to deal with language and cultural barriers, besides having to cope with greater levels of uncertainty and risk (Teele, 2014).

Furthermore, because development RCTs are implemented by mostly US-based researchers in developing countries, the distinctions between economic knowledge-making and development policy governance is particularly tricky to manage. As we will see in the third section, this is because development RCTs are more than simple hypothesis-testing instruments for economic theory, they are also used to redistribute social resources, to measure the impact of development projects, and to propose directions for future foreign aid interventions (Rottenburg et al., 2015). In this process, the line between what counts as experimentation and what counts as a development project is constantly in flux, posing similar ethical questions as processes used for recruiting human subjects for global clinical trials (Petryna, 2009; Rottenburg, 2009), related to whether development RCTs exploit or aim to help the global poor.

In sum, development RCTs combine aspects of both forms of previous experimentation in economics, but applied to a novel territory of intervention, that is, developing countries. On the one hand, development RCTs retain some characteristics of lab science, such as control groups and the aim to identify the causal effect of interventions. On the other hand, they are decisively in the field, as they are conducted "in the messiness" of everyday life of developing countries, "where borders cannot be effectively policed" (Henke, 2000, p. 484). Building on Gieryn (2006, p. 32), therefore, similar to other field sciences, development RCTs gain legitimacy by "preserv[ing] and draw[ing] simultaneously—and in a complementary way—the assumed distinctive virtues of both lab and field." How do *randomistas* manage to do so?

3 | METHODS AND DATA

To answer this question, I use two analytical strategies. First, I build on a dataset constructed as part of a larger project, in which I compiled a random sample of 63 RCTs done by the Abdul Latif Jameel Poverty Action Lab at MIT, or JPAL, led by the charismatic scholar-activist Esther Duflo. My analytical sample was defined on January 13, 2016. Of the 625 RCTs that were listed in J-PAL's library that day, I excluded 100 studies that were not conducted in developing countries. From 525 RCTs, I randomly selected 100 RCTs to be analysed. I then excluded all RCTs that were still ongoing or for which I could not identify a corresponding publication, arriving at a final sample of 63 RCTs. For each RCT publication, I coded information regarding their study design, the authors, and relevant information about implementing and funding partners (see the Appendix for descriptive statistics of the sample).

Second, my account relies on 4 years of participant observation with *randomistas*. During this period, I participated in two RCTs related to microfinance in Peru (2007), and one RCT in the field of financial education in

Brazil (2010–2012). For the purposes of this paper, I complemented this ethnographic data with an analysis of the controversy regarding development RCTs that appears in academic and policy debates. The publications analysed include academic articles, blog posts, and public interviews given by *randomistas* and international policy actors. It is to the analysis of this data that I now turn.

4 | THE MAKING OF DEVELOPMENT RCTS: ACTORS, PROCESSES, AND CONTROVERSIES

In her TED Talk, Esther Duflo (2010) explained how *randomistas* would revolutionize the international development field:

> *It's not the Middle Ages anymore, it's the 21st century. And in the 20th century, RCTs have revolutionized medicine by allowing us to distinguish between drugs that work and drugs that don't work. You can do the same randomized controlled trial for social policy. You can put social innovation to the same rigorous, scientific tests that we use for drugs.*

As has been extensively documented, however, the "revolution" that RCTs brought to the medical field depended on a contentious process that required political and organizational efforts to convince the multiple actors involved of the possibilities of adopting the experimental method with human subjects (Carpenter, 2010). The same is true for development RCTs. Implementing the experimental method in field conditions to answer international development questions is a huge task that involves multiple actors and resources.

In this section, I build on both my RCT sample and on my ethnographic data to describe how economists implement field experiments in developing countries; how they enrol relevant actors into field experiments; and the chain of transcriptions required to generate data from poor individuals' behavior in the field that will then be published in academic journals. The findings in this section unveil the type of stakeholder enrolment and organizational efforts that enable *randomistas* to establish the idea that developing countries are appropriate field-sites for scientific analysis—that is, sites that are not contaminated by geopolitical interests or bureaucratic politics. Through this assessment, I also demonstrate how the boundary work done to differentiate development research and practice allows *randomistas* to productively dismiss failures in field experiments as ideological or bureaucratic problems.

4.1 | The RCT network: A diverse set of actors

While much attention has been paid to *randomistas*, who are the public faces of RCTs and arguably the most influential actors in the network (Ogden, 2016), implementing any given development RCT requires collaboration among a number of actors: the research team, the fieldwork team, survey firms, policy beneficiaries, funding agencies, and implementing partners. Below, I briefly describe each of these in order to highlight the diverse set of actors involved in development RCTs, as well as the internal negotiations that happen among these actors to make these experiments possible.

1. Research Team: The implementation of an RCT starts with the Research Team, based in Poverty Labs within economic departments, such as JPAL at MIT. In my analytical sample, 93% of these researchers had received their PhDs in Economics, and 7% graduated in Political Science. *Randomistas* have academic and administrative roles: they formulate the research question, design the experiment, analyse the data, and publish the academic papers, but they also have a key role in convincing the policy community of

the relevance of their work, in building the reputation of their labs, and in negotiating with funding agencies.

2. Field Team: Although the Research Team formulates the experimental design, implementing an RCT requires that an extended group of professionals work directly in developing countries, close to the project location. These individuals have a different profile than *randomistas*. From the 123 local staff that appear in my sample, 45% had graduate degrees in Economics and Public Policy, but many instead had graduate degrees in Development Studies (20%) and even in the Humanities (12%). The Field Team has a diverse range of tasks—from explaining the experimental methodology to partners implementing the policy and asserting the quality of the experiment, to hiring survey firms and solving any unexpected problems in the field. As a *randomista* explained to me, they are considered "the voice and eyes of the researchers" in the project site, and play a key role in controlling the quality of the experiment. Below, I will show how these actors also have great discretionary power in the transcription process involved in RCTs.
3. Implementing Partners: These are the organizations whose policies will be evaluated by the Research Team. Partner organizations can be divided into high-level decision-makers and their staff (which I treat as a separate actor). The vast majority of RCTs are conducted with NGOs: in my analytical sample, for example, 75% of implementing partners are NGOs or for-profit organizations involved in microfinance projects together with local NGOs. In my ethnographic work, I observed that these NGOs are led by a highly educated group with extensive international experience, which came either from their relationship with foreign aid agencies or professional training. The role of these policy managers in RCT implementation is to establish the institutional partnership with Poverty Labs, provide financial and infrastructural resources, and disseminate results together with *randomistas*.
4. NGO Staff: This group is responsible for the day-to-day work of development organizations. Staff members include teachers, nurses, and micro-credit agents, among others. While I did not have information about these actors in my RCT sample, during my fieldwork, I could examine the key features of their profile: they have either secondary education or a BA degree in a less prestigious field of study; they usually do not speak English, but they have a great deal of tacit knowledge regarding the policy being implemented. NGOs' staff are the ones that have their daily activities most affected by the experiment: they modify their practices to respect the treatment and control group division, report back to the Field Team and the high-level staff about their activities, and make the connection between the survey firm and policy beneficiaries, while also being directed evaluated.
5. Policy Beneficiaries: These are the individuals, or the local population, who are affected by the policy being tested and will respond to the survey questionnaires. *Randomistas* refer to them as "The Poor" (a broad category that involves a multiplicity of groups) whose behavior they are trying to understand and shape. Results from field experiments depend on their willingness to answer survey questionnaires properly, but they have no influence over the design of the RCTs. Yet, all administrative data available of their behavior and socioeconomic background is closely monitored to assess the impact of development policies.
6. Survey Firms: In order to obtain data from development RCTs, a crucial part of the process is hiring and training either a local survey firm or independent surveyors. In my fieldwork, I noticed that surveyors had a university degree and experience with survey implementation, but no international work experience or English proficiency. Surveyors are trained to implement the questionnaires with the policy beneficiaries, are taught the basics of the experimental method, and have their activities closely supervised by the Field Team.
7. Funding Agencies: Poverty Labs count on a diverse portfolio of agencies to fund their activities. In my sample, 34% of funding came from multi-lateral institutions (e.g., World Bank and USAID), 40% from philanthropic foundations (e.g., Bill and Melinda Gates Foundation), and the rest from a mix of local sources and country donors. Their role in implementation is to provide the financial and infra-structural resources, but Funding Agencies can also have an impact on the type of projects that will be evaluated. Individuals in these agencies have a similar profile to the policy-makers: elite academic training and vast international experience.

4.2 | Chains of transcriptions: The process of implementing an RCT

The actors in the RCT network are situated in distinct geographical locations, and they influence field experiments with different weights and strategies. Perhaps the most powerful actors in the network are located in academic departments in the US or Europe, where the Research Team and Funding Agencies make decisions regarding the design and costs of RCTs and decide which NGOs to partner with at the local level. The immense infrastructure required to implement RCTs, however, is situated in remote areas of developing countries. Although NGOs, Surveyors, and Policy Beneficiaries have little influence on the design of the experiments, without their active cooperation, the RCT could not be implemented.

Contrary to the transcription of natural objects (Latour, 1999), transcribing the behavior of poor individuals depends greatly on the ability of researchers to govern the web of social relations that exist in the field. To do so, the transcription process starts with an intense communication exchange between the Research and Field Teams to decide on the best experimental design possible to evaluate the NGO's policy. Both teams use available administrative data to estimate the minimum size and design of the experiment that can simultaneously guarantee statistical power (hence securing the robustness of the results), while sounding politically feasible.

After a decision is reached about the experimental design at the Poverty Lab level, another round of analysis and negotiation happen between the Field Team, high-level officials, and staff from NGOs. This is when the preferred experimental design by the researchers gets a reality check from policy managers, in what can be a very contentious negotiation process. In my fieldwork in Brazil, for example, one teachers' union threatened to strike against what they considered an overly ambitious RCT size. In response, the Research Team substantially reduced the experimental design to secure teachers' buy-in and the continuity of the experiment. Similar to this case, it is common for the RCT strategy to be simplified based on constraints of funding, data, ethical and political concerns, infrastructure, and operations. For a development RCT to happen, policy and academic sides have to compromise and agree on a final experiment design.

With an experiment strategy in hand, the Field Team is then responsible for hiring a survey firm that will apply the questionnaires and do the data-entry of the surveys, as well as training the street-level bureaucracies implementing the policies to adapt their practices to fit the experiment design. At this stage, again, local bureaucracies often react against attempts to substantially change their daily practices in the name of the experiment, prompting the Field Team to continuously remind them about the importance of securing the quality of the RCT and the surveillance mechanisms they will use to ensure that the NGO staff does so. The training period is thus key to avoid that "NGOs boycott the RCT" and that they "appreciate the value of our experiment," as pointed out by two Field Team members in Peru.

Following this training period, an initial questionnaire is implemented with all the beneficiaries from control and treatment groups before the new policies being tested are put into practice. It is during the implementation of this baseline survey that the journey of the RCT data from the field to publications starts. A first, non-trivial, step is finding the individuals that will be interviewed—the ones from the treatment group are relatively easier to find because surveyors can count on the NGOs' administrative apparatus to schedule the interviews. Interviewing individuals in the control group, however, is harder because these individuals will not benefit from the policy, they commonly live in places without formal addresses, and they have to voluntarily agree to serve as research subjects (Rayzberg, 2019a).

After individuals from control and treatment groups are found, the actual paper survey can then be implemented. Usually, understanding how "The Poor" behaves means asking individuals to wait in line to be interviewed and to answer a 10- to 50-page-long questionnaire. Interviews last from 30 min to 2 hr. They are done in the local language of the village where the experiment will take place, and the interview schedule consists mostly of behavioral questions, which are frequently hypothetical questions, such as the examples below:

Determining financial literacy	*Determining aversion to risk*
Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, would you be able to buy more than, exactly the same as, or less than today with the money in this account?	*Suppose you had to choose between the following two options:* *A. You receive 30 dollars with certainty.* *B. A coin toss: if the outcome is heads, you receive 100 dollars; if it is tails, you receive nothing. Do you understand the two options? [Check for understanding]. Which would you prefer?*
a) More than today;	*a) 30 dollars with certainty;*
b) Exactly the same as today;	*b) 100/0 dollars coin toss;*
c) Less than today.	*c) Don't know.*

Through answers to one or two questions like the examples above, researchers will then create an index measuring, for example, "Financial Literacy Levels" of respondents, as well as "Aversion to Risk" indexes. Besides hypothetical behavioral questions, questionnaires have socioeconomic and policy implementation questions to guarantee that the control and treatment group divisions were respected.

After the baseline survey is finished, the development intervention begins. The time distance between this first survey and the final survey, when the results of the policies will be assessed, varies, ranging from 3 months to up to 2 years. During this period, NGO staff are expected to respect the division between control and treatment groups, and the fieldwork team is constantly monitoring if they do so. The forms of surveillance of the quality of the experiment are multiple and key to the success of the RCTs: they range from surprise visits to different project sites and monitoring of administrative data of the NGO, to meetings with beneficiaries and policy officials, and a great deal of face-to-face conversations to remind street-level staff of the importance of respecting the design of RCTs.

Finally, after the policy being tested ends, a second survey is administered with the same questionnaire, and the data-entry of all the paper surveys begins. This data is intensely "cleaned" by the Field Team, then sent to universities where the analysis and eventual publications will be carried out. At the end of this data-cleaning process, *randomistas* receive a database file containing the behavior of "The Poor." Even at this point, when the data has already been completely decontextualized and is ready to be analyzed, there will still be many communications between *randomistas* and the Field Team to explain non-intuitive results and signs of control group contamination before researchers agree on a final analysis. Even if individuals in the Field Team are the ones with the lower credentials in the academic side of the RCT network, they have a lot of discretionary power to explain what happened in the field for the authors of the final papers.

Hopefully at this point it is clear that the process of implementing development RCTs is long and demands coordination efforts between actors based in different continents who speak different languages and have different cultural practices, in addition to demanding a great deal of financial and material resources. In each of the steps that I described, adaptations and pragmatic decisions to deal with shortcomings are made. Moreover, considering the long time that RCTs take from their initial formulation until the data from the final survey is collected, the experiments have to deal with a lot of discontinuity and turnover from NGOs' staff, survey firms, and even in the Field Team, which contributes to making the chains of transcriptions from the field to universities even more complex and attempts to control this process even more ardent. Similar to what Rosengarten and Savransky (2019) describe about medical RCTs, these adaptations render evidence produced by development RCTs incredibly situated and dependent on the specific relational dynamics that characterize their implementation.

4.3 | Controversies: Controlling the control group and randomizing social benefits

For all the apparatus described above to function, the biggest challenges for both researchers and implementing partners are randomization and the quality of "no-treatment" control groups. It is only by solving these two issues that *randomistas* guarantee the scientific status of their work and are thus able to portray an image of their field experiments as free from politics or bureaucratic interference—that is, as different from the image of development work as inherently corrupt and inefficient, as suggested by some prominent economists (Easterly, 2007). Yet, considering that they are working with development projects, there is nothing trivial about solving these challenges.

It is no surprise that the creation of a randomly selected "no-treatment" control group with foreign aid beneficiaries is politically and ethically controversial (de Souza Leao & Eyal, 2019). Whenever the development intervention involves the distribution of resources and services, assigning individuals to control groups inevitably raises strong resistance from the development community. In fact, randomized assignment can easily be considered illegal (Glennerster, 2015), unless the development intervention can be framed as a non-entitlement, or if it can be framed as merely a "nudge" meant to overcome behavioral obstacles rather than a form of assistance (Berndt, 2015). How can bureaucrats, NGO staff, or politicians justify giving financial resources and social services to some people but not to others who need just as much? For this reason, *randomistas* have to show that their random assignment was not affected by political or bureaucratic considerations, and thus have to deal with the resulting constraints associated with providing benefits solely for one part of the target population (and withholding benefits from others equally in need) in the name of the scientificity of their experiments.

Moreover, in cases when randomization does happen, welfare regulations and political pressures create the risk of "substitution bias"—when individuals in the control group have good substitutes for the tested policy—which could contaminate the quality of experimental and control groups, and hence undermine the trust in the RCT findings (de Souza Leao & Eyal, 2019; Heckman, Hohmann, Smith, & Khoo, 2000). This means that the scientific success of the RCT network described above is contingent on *randomistas'* ability to show that participants in the experiment did not benefit from other social policies that could interfere with the results—something that would be impossible in richer countries, but that is feasible in remote areas of developing countries.

The tasks of randomizing social benefits and controlling the control group are made both more manageable and more challenging if we consider the normative and regulatory environment that characterizes the current international development field (Watkins, Swidler, & Hannan, 2012). First, in response to the widespread perception that foreign aid was ineffective and corrupt, the 2000s have been characterized as a period of traditional "aid fatigue," and by the entry of a new set of actors, mainly private foundations and NGOs, to the development field (Easterly, 2007; Krause, 2014). As mentioned above, the fact that researchers are now working with NGOs, rather than local governments as was done before, means that they can minimize the political and ethical problems posed by randomization (de Souza Leao & Eyal, 2019).

Second, the privatization of foreign aid has also altered the *type* of aid that is disbursed and under what conditions, and has created new accountability struggles for these new actors (Krause, 2014). Considering that aid typically flows from foreign donors to global NGOs, who are the ones responsible for selecting local partners who will then implement projects in small villages where "development" is supposed to happen, many authors have pointed to the principal–agent problem that characterizes the current aid chain (Swidler & Watkins, 2009). On the one hand, this long funding chain, combined with the fact that NGOs operate in unfamiliar cultural and political terrains marked by "the loss of hope in development" (Krause, 2014, p. 42), results in an aid environment characterized by strong attempts at control by donors and by a focus on measuring aid effectiveness. On the other hand, there is great suspicion and criticism of whom the new private donors and *randomistas* are accountable to, turning the rhetoric around experimenting on the poor of the Global South into a target of criticism:

> *Donors increasingly want to see more impact for their money… Some go so far as to insist that development interventions should be subjected to the same kind of randomised control trials used in medicine,*

> *with "treatment" groups assessed against control groups… But truly random sampling with blinded subjects is almost impossible in human communities without creating scenarios so abstract as to tell us little about the real world. And trials are expensive to carry out, and fraught with ethical challenge… People of the south deserve better. (Op-ed signed by 15 leading economists in* The Guardian[1]*)*

In sum, *randomistas* deal with two types of controversies to successfully portray development RCTs as scientific and free from politics or other ideological struggles that characterize the foreign aid field. On the academic side, researchers have to convincingly show that assignment to treatment and control groups was random, without political interference, and that throughout the experimental period the control group was not contaminated. However, the same characteristics that make development RCTs scientifically rigorous are the ones that make them so politically controversial. This is because, on the policy side, *randomistas* have to convince other actors of the importance of randomizing the distribution of social benefits and controlling the control group, while dealing with the fear of corruption of development projects and with criticisms of the advisability of foreign aid. The latter issue is particularly salient in debates about post-colonialism and the role that foreign funders have in developing countries (Li, 2007; Tilley, 2011).

5 | BOUNDARY WORK BETWEEN TESTING AND DOING DEVELOPMENT

To deal with these controversies, *randomistas* build on the strategic ambiguity of what is on trial to distance themselves from the politics of foreign aid, while being close enough to the field to dictate what works in international development and for other development actors to find partnering with Poverty Labs attractive. Hence, while development RCTs are conceptualized as a means of finding the causal impact of a policy, they are simultaneously touted as serving to test economic theories, making it purposively unclear what is being tested—is it a development intervention, an economic theory, or both? The excerpt below is typical of the strategic ambiguity adopted by *randomistas*:

> Can a RCT tell us not just whether an intervention worked, but also how and why? *When designed and implemented correctly, RCTs can not only tell us whether an intervention was effective, but also answer a number of other policy-relevant questions… However, as with any single study, a RCT is just one piece in a larger puzzle. By combining the results of one or more RCTs with economic theory, descriptive evidence, and local knowledge, we can gain a richer understanding of an intervention's impact. (JPAL, 2017)*

Furthermore, *randomistas* avoid the political debate about development and foreign aid by testing development economic theories that are quite simple, such as in the RCT example introduced earlier: "If you give deworming pills to school children, they get sick less frequently, and go to school more often" (see also Abdelghafour, 2017). Even when *randomistas* limit themselves to these small, short-term questions, they deem each trial as answering one part of broader questions, which can guarantee their contribution to development economic theory. In this way, *randomistas* connect themselves back to the discipline of economics, not to program implementation or the more complicated politics of foreign aid (de Souza Leao & Eyal, 2019).

To put it more generally, *randomistas* resolved the political problem posed by randomization with foreign aid beneficiaries by maintaining a productive ambiguity so that RCTs mean different things to different constituencies at different times. Not only do field experiments equivocate between exercises in theory-building and solutions to policy problems, but as Rayzberg (2019a) shows, *randomistas* employ multiple devices to frame the RCT as an ambiguous object, partly a development intervention and partly a test, in order to overcome political and practical obstacles to randomization. If the RCT is an intervention, local inhabitants and NGO staff would be keen to participate, but would not consider it legitimate to arbitrarily deny the benefits to some participants in the control group.

If the RCT is testing an economic theory, assignment to the control group would be legitimate, but people would be much less eager to participate. The various different designs of development RCTs, Rayzberg (2019a) suggests, can be understood as framings meant to contain and manage this problem, to entangle intervention and test together so as to be able to recruit participants, but then also to disentangle them, so randomization is possible.

As a result, it is common for development RCTs to have a design in which the control group receives some kind of treatment—either a reduced version or the status quo. The lesser treatment is a boundary object. The researchers consider it the null condition of the theory-building test; the participants consider it an intervention. This is also why experimental phase-in design is so popular.[2] Phase-in is a temporal framing device (Rayzberg, 2019a). During the first-time frame, the control group receives no treatment; but during the second time frame it receives the same treatment as the experimental group. So intervention and testing are disentangled in the present, but entangled in the future. Phase-in makes the perceived unfairness of being in a control group much more manageable, since participants are promised to receive social benefits in the near future.

5.1 | When the boundary work fails

The mechanisms illustrated by Rayzberg (2019a) are important for understanding the boundary work involved in making development RCTs successful. Similar to what Gieryn (1983) explained about the continuous need to demarcate boundaries between science and varieties of non-sciences in order to establish scientific authority, *randomistas* construct boundaries between their economic research trials and simple program evaluation in order to secure their space in the two worlds. This boundary work, however, is ongoing and historically changing, and many times *randomistas* are not able to maintain the productive ambiguity between research and development intervention. My analysis found two main reasons why the boundary work typically fails.

The first reason why *randomistas* are unable to maintain the boundary between research and intervention is that policy beneficiaries sometimes discover that they are in control groups and revolt against the experiment. By adopting the phase-in design mentioned above, *randomistas* are usually able to bypass this problem by promising that individuals will benefit from the development project in the future, but it is common for participants to confront researchers about their status as members of the control group, even if they are promised future gains. As I often heard in fieldwork: “Being in the control group rarely goes unnoticed.” For example, in my observations of microcredit RCTs in Peru, individuals in the control group complained multiple times to the NGO that surveyors asked questions about their financial behavior in order to limit their chances of receiving credit in the future. While surveyors were employed by the Poverty Lab, control group individuals conflated their work with the work of the microcredit NGO.

In cases similar to this one, participants blurred the boundaries between research and intervention by holding *randomistas* responsible for both their own and the NGO’s actions, while the Field Team continuously insisted on their distinctiveness. In this case, the Field Team demarcated the boundaries between their role as researchers and the role of the NGO. They were asking questions about the financial behavior of families to learn how “The Poor” made financial decisions, not to limit or facilitate access to microcredit. Yet, while the distinction between research and intervention was important for the Field Team, for “research participants this difference is often arbitrary, irrelevant, or entirely meaningless”, as pointed out by Rayzberg (2019a, p. 389).

Second, and relatedly, there are cases in which high-level policy officials push back against the interests of *randomistas*. This opposition may be related to the perception that Poverty Labs’ research “draws its motivation from academic concerns that overlap imperfectly with the issues that matter to development practitioners” (Ravallion, 2009, p. 48); or to the specific politics of the programs that *randomistas* choose to evaluate. When this happens, *randomistas* are rarely able to maintain the productive ambiguity between research and intervention and face backlash regarding the political goals of their tests and how connected they are to broader ideological and political agendas happening in the foreign aid field. In other words, their field becomes politicized or contaminated, diminishing their ability to claim its scientific status. A recent battle between UK-based Action Aid and *randomistas* can serve to illustrate this type of resistance.

In this RCT, three US-based economists partnered with Liberia's federal government to test the effectiveness of a public-private partnership to introduce what they referred to as "American-style charter schools or the UK's academies to Liberia's underperforming education system" (Romero, Sandefur, & Sandhotz, 2017). In order to implement the RCT, however, *randomistas* insisted that the Partnership Schools for Liberia (PSL) program be first tested as a pilot in order for experimental evidence to be generated. As a reaction, Action Aid released a Request for Proposal for qualitative researchers to closely observe the PSL pilot (Edwards, 2017). Action Aid saw many problems with this RCT: (a) it involved an unrealistic financial investment in schools that would not be reproducible at a larger scale; (b) it aimed to weaken public education in favor of *specific* private partners; and (c) it did not answer the right questions:

> *[We have] concerns about whether it is feasible for the RCT, however rigorous, to isolate the added value that arises from the involvement of private providers. How will the government of Liberia or anyone else learn anything, other than the already evident truth that if you spend more money per student and have smaller class sizes you will get better results? (Archer, 2016)*

While defending the pilot, *randomistas* got entangled in the politics of post-war Liberia, choosing to publicly defend Liberia's president and minister of education for their commitment to piloting the new educational policy. Similar to the Field Team strategy presented above, *randomistas* insisted on the strong boundaries and differences between their research and the politics of the PSL:

> *We also agree with Action Aid and Education International on more substantive matters… The key difference is that we do not read these points as criticisms of our study. Rather, many of the points below are concerns over the program itself, many of which we share… We would reframe these concerns as hypotheses that we have explicitly designed the randomized evaluation to test. (Romero et al., 2017)*

Yet, even with their attempt to separate their experiment from the politics of the pilot, the public backlash has been huge, and researchers are now accused of having a hidden agenda of privatization of educational systems around Africa. Indeed, Education International—a partner organization of Action Aid—included the PSL pilot in its global campaign to "oppose privatisation and commercialisation of public education" (EI, 2017), making the controversy visible to an even wider audience. As aptly put by a World Bank official[3]: "The new Liberia school experiment is destined to be a Rorschach test of which side of the private education debate you sit on." Whether this was the intention of researchers or not, their scientific neutrality has been put into question by parts of the development community.

By characterizing their critics as doing "advocacy not research," however, the three authors did not see their test as politicized, claiming instead that they were collaborating for an "ongoing debate that [was] increasingly disciplined by facts" (Romero et al., 2017). Put differently, when the boundary work failed, *randomistas* quickly reinforced the boundaries between research and intervention, and put on their "scientist hats," dismissing development actors as driven by "conviction" or, as Esther Duflo, has claimed multiple times, as being driven by the "three I's—Ideology, Ignorance, and Inertia" (Duflo, 2017, p. 9).

6 | CONCLUSION

In this paper, I have argued that the key element for understanding the current success of development RCTs is unveiling how *randomistas* both draw and blur the boundaries between development practice and economic research in order to implement field experiments. To do so, I presented the massive organizational effort and the diverse set of actors involved in the making of RCTs and the continuous boundary work that *randomistas* do to differentiate the purposes and merits of testing development projects from doing them, as a way to bypass the problems presented by adopting the experimental method with foreign aid beneficiaries in poor countries. In conclusion, I address two implications of this research for the study of testing and tests.

First, continuing a tradition in science and technology studies, my research draws attention to the ongoing negotiations between the desire to control and the need for cooperation that characterize the implementation of field experiments (Henke, 2000; Kohler & Vetter, 2016). To this end, the case of development RCTs demonstrates that successful field experimentation depends on the researcher's ability to productively manage the ambiguity of what is being tested. Yet, contrary to many field sciences, in which scientific authority results from an intense familiarity with certain places (Gieryn, 2006), in development RCTs, managing this ambiguity means that *randomistas* distance themselves from the broader history, politics, and particularities of their object of study, purposively dissociating developing countries from the broader dynamics of the foreign aid industry. As a consequence, development RCTs differ greatly from the "heroic interventions of modernity, based on the narrative of progress" that often characterized foreign aid, focusing instead on diffuse, "evidence-based melioristic interventions" that are portrayed as behavioral in nature (Berndt, 2015; Rottenburg et al., 2015, p. 10).

Second, my findings connect to debates about the performativity of tests (MacKenzie et al., 2007). One of the arguments that I put forward in this paper is that the credibility of RCTs to the scientific community is contingent on the belief that field experiments were not influenced by material, ideological, or political interests. Yet, similar to other types of tests, my research shows that the attempts to control developing countries make development RCTs extremely performative, since their success depends on development projects adapting their operations to fit experimental procedures. Indeed, as the dissemination of RCTs takes place, research and policy designs increasingly start to look like textbook RCTs. The UK government, for instance, recently published a policy paper—"Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials" (Haynes, Service, Goldacre, & Torgerson, 2012), suggesting that all government policies be designed with an RCT evaluation in mind.

At a time when testing and global policy standards are diffusing with higher speed than before, my findings are also relevant to assess the potentially exclusionary nature of top-down control over acceptable forms of evidence. As the case of development RCTs rests clear, the imposition of evidence hierarchies in international development and its consequences for the distribution of global aid are based on the premise that development projects can and should be modified for testing purposes. As I have demonstrated, however, transforming developing countries into appropriate "fields" for experimentation prompts resistance, misunderstandings, and uneven negotiations that often exclude the voice and interests of local beneficiary populations. Moving the conversation forward, more qualitative research on the views of these actors is needed to envision the possibilities of creating alternative models of evaluation, as well as more democratic methods of transferring resources to countries from the Global South.

ACKNOWLEDGEMENTS

The author wants to thank Diego Povoas and XingJian Li for their work in this project, as well as Gil Eyal, Margarita Rayzberg, and participants of the Put to the Test workshop at Warwick University for their helpful comments and suggestions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

NOTES

[1] "Buzzwords and Tortuous Impact Studies Won't Fix a Broken Aid System," *The Guardian online*, July 16, 2018. Available at https://www.theguardian.com/global-development/2018/jul/16/buzzwords-crazes-broken-aid-system-poverty [Accessed on November 2, 2018].

[2] In my sample, in 35% of studies the control group received either the status quo (i.e., the development project) or a slightly reduced version of it. In 40% of studies, researchers adopted the phase-in design, in which the control group receives the development policy in a later period. In the remaining 25% of studies, the control group did not receive any type of treatment.

[3] Matt Collin in Twitter, September 8, 2017. Available at https://twitter.com/aidthoughts/status/906181438390902784 [Accessed on November 2, 2018].

REFERENCES

Abdelghafour, N. (2017). Randomized controlled experiments to end poverty. *Anthropologie & Développement*, *46–47*, 235–262.

Archer, D. (2016). *The challenges of education reform and privatization in Liberia*. Retrieved from http://www.actionaid.org/2016/12/challenges-education-reform-and-privatisation-liberia.

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

Berndt, C. (2015). Behavioral economics, experimentalism and the marketization of development. *Economy and Society*, *44*(4), 567–591.

Carpenter, D. (2010). *Reputation and power—Organizational image and pharmaceutical regulation at the FDA*. Princeton, NJ: Princeton University Press.

De Souza Leão, L., & Eyal, G. (2019). The rise of Randomized Controlled Trials (RCTs) in international development in historical perspective. *Theory and Society*, *48*(3), 383–418.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*(2), 424–455.

Deaton, A., & Cartwright, N. (2016). Understanding and misunderstanding randomized controlled trials. *NBER Working Paper Series*, No 22595.

Duflo, E. (2003). Poor, but Rational? *MIT Working Paper*, *747*.

Duflo, E. (2010). Social experiments to fight poverty. *TED Conference*, Retrieved from https://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty.

Duflo, E. (2017). The economist as plumber. *American Economics Review: Papers & Proceedings*, *107*(5), 1–26.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. CEPR Discussion Papers, No 6059.

Easterly, W. (2007). *The white's men burden: Why the west's efforts to aid the rest have done so much ill and so little good*. New York: Penguin USA.

Edwards, S. (2017). *Funding struggle: Can Liberia's controversial privately run schools pilot continue?* Retrieved from https://www.devex.com/news/funding-struggle-can-liberia-s-controversial-privately-run-schools-pilot-continue-91061.

EI. (2017). *"Development cooperation" and "unite for quality education campaign"*. Retrieved from https://ei-ie.org/en/detail_page/4641/development-cooperation.

Escobar, A. (2012). *Encountering development: The making and unmaking of the third world*. Princeton, NJ: Princeton University Press.

Eyal, G. (2013). For a sociology of expertise: The social origins of the autism epidemic. *American Journal of Sociology*, *118*(4), 863–907.

Ferguson, J. (1990). *The anti-politics machine: Development, depoliticization, and bureaucratic power in Lesotho*. Cambridge: Cambridge University Press.

Gieryn, T. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review*, *48*(6), 781–795.

Gieryn, T. (2006). City as truth-spot: Laboratories and field-sites in urban studies. *Social Studies of Science*, *36*(1), 5–38.

Glennerster, R. (2015). *So you want to do an RCT with a government: Things you should know*. Retrieved from http://runningres.com/blog/2015/12/9/so-you-want-to-do-an-rct-with-a-government-things-you-should-know.

Guala, F. (2007). How to do things with experimental economics. In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets?* Princeton, NJ: Princeton University Press.

Guggenheim, M. (2012). Laboratizing and de-laboratizing the world: Changing sociological concepts for places of knowledge production. *History of the Human Sciences*, *25*(1), 99–118.

Harrison, G. (2011). Randomization and its discontents. *Journal of African Economies*, *20*(4), 626–652.

Haynes, L., Service, O., Goldacre, B, & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Retrieved from https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials.

Heckman, J., Hohmann, N., Smith, J., & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics*, *115*(2), 651–694.

Henke, C. (2000). Making a place for science: The field trial. *Social Studies of Science*, *30*(4), 483–511.

JPAL. (2017). *Jameel poverty action lab website*. Retrieved from www.poverty-action.org.

Kabeer, N. (2019). Randomized control trials and qualitative evaluations of a multifaceted programme for women in extreme poverty: Empirical findings and methodological reflections. *Journal of Human Development and Capabilities*, *20*(2), 197–217.

Kohler, R., & Vetter, J. (2016). The field. In B. Lightman (Ed.), *A companion to the history of science*. Chichester: John Wiley & Sons.

Krause, M. (2014). *The good project: Humanitarian relief NGOs and the fragmentation of reason*. Chicago, IL: Chicago University Press.

Latour, B. (1999). *Pandora's hope: Essays on the reality of scientific studies*. Cambridge, MA: Harvard University Press.

Li, T. (2007). *The will to improve: Governmentality, development, and the practice of politics*. Durham, NC: Duke University Press.

MacKenzie, D., Muniesa, F., & Siu, L. (2007). *Do economists make markets? On the performativity of economics*. Princeton, NJ: Princeton University Press.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.

Ogden, T. (2016). *Experimental conversations: Perspectives on randomized trials in development economics*. Cambridge, MA: MIT Press.

Petryna, A. (2009). *When experiments travel: Clinical trials and the global search for human subjects*. Princeton, NJ: Princeton University Press.

Ravallion, M. (2009). Evaluation in the practice of development. *The World Bank Research Observer*, *24*(1), 29–53.

Rayzberg, M. (2019a). Fairness in the field: The ethics of resource allocation in randomized controlled field experiments. *Science, Technology and Human Values*, *44*(3), 371–398.

Rayzberg, M. (2019b). *Controlling the field: Experimental social science and politics of evidence in international development* (PhD dissertation). Northwestern University, Sociology Department, Chicago, IL.

Riecken, H., & Boruch, R. (1975). *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.

Rogers-Dillon, R. (2004). *The welfare experiments: Politics and policy evaluation*. Stanford, CA: Stanford University Press.

Romero, M., Sandefur, J., & Sandholtz, W. (2017). *Will an RCT change anyone's mind? Should it?* Retrieved from https://www.cgdev.org/blog/will-rct-change-anyones-mind-should-it.

Rosengarten, M., & Savransky, M. (2019). A careful biomedicine? Generalization and abstraction in RCTs. *Critical Public Health*, *29*(2), 181–191.

Rottenburg, R. (2009). Social and public experiments and new figurations of science and politics in postcolonial Africa. *Postcolonial Studies*, *12*(4), 423–440.

Rottenburg, R., Merry, S., Park, S., & Mugler, J. (Eds). (2015). *The world of indicators: The making of governmental knowledge through quantification*. Cambridge: Cambridge University Press.

Swidler, A., & Watkins, S. (2009). "Teach a man to fish": The sustainability doctrine and its social consequences. *World Development*, *37*(7), 1182–1196.

Teele, D. (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. New Haven, CT: Yale University Press.

Tilley, H. (2011). *Africa as a living laboratory: Empire, development, and the problem of scientific knowledge, 1870–1950*. Chicago, IL: Chicago University Press.

Watkins, S., Swidler, A., & Hannan, T. (2012). Outsourcing social transformation: Development NGOs as organizations. *Annual Review of Sociology*, *38*, 285–315.

How to cite this article: de Souza Leão L. What's on trial? The making of field experiments in international development. *Br J Sociol*. 2020;71:444–459. https://doi.org/10.1111/1468-4446.12723

APPENDIX

Analytical Sample

TABLE A1 Descriptive statistics

Total (*n* = 63)	
Countries	23 different countries
Average size of RCT	8,660 individuals
Median size of RCT	2,156 individuals
Average duration	12 months
RCT topic: Finance	38%
RCT topic: Education	24%
RCT topic: Health	22%