Neuwinger, Malte

Article — Published Version

What makes randomized controlled trials so successful —for now? Or, on the consonances, compromises, and contradictions of a global interstitial field

Theory and Society

Suggested Citation: Neuwinger, Malte (2024) : What makes randomized controlled trials so successful —for now? Or, on the consonances, compromises, and contradictions of a global interstitial field, Theory and Society, ISSN 1573-7853, Springer Netherlands, Dordrecht, Vol. 53, Iss. 5, pp. 1213-1244, https://doi.org/10.1007/s11186-024-09564-5

This Version is available at:
https://hdl.handle.net/10419/315623

Theory and Society (2024) 53:1213–1244
https://doi.org/10.1007/s11186-024-09564-5

What makes randomized controlled trials so successful—for now? Or, on the consonances, compromises, and contradictions of a global interstitial field

Malte Neuwinger[1]

Accepted: 26 May 2024 / Published online: 20 June 2024

Abstract
Randomized controlled trials (RCTs) are a major success story, promising to improve science and policy. Despite some controversy, RCTs have spread toward Northern and Southern countries since the early 2000s. How so? Synthesizing previous research on this question, this article argues that favorable institutional conditions turned RCTs into "hinges" between the fields of science, politics, and business. Shifts toward behavioral economics, New Public Management, and evidence-based philanthropic giving led to a cross-fertilization among efforts in rich and poor countries, involving states, international organizations, NGOs, researchers, and philanthropic foundations. This confluence of favorable institutional conditions and savvy social actors established a "global interstitial field" inside which support for RCTs has developed an unprecedented scope, influence, operational capacity, and professional payoff. However, the article further argues that the hinges holding together this global interstitial field are "squeaky" at best. Because actors inherit the *illusio* of their respective fields of origin—their central incentives and stakes—the interstitial field produces constant goal conflicts. Cooperation between academics and practitioners turns out to be plagued by tensions and contradictions. Based on this analysis, the article concludes that the global field of RCT support will probably differentiate into its constituent parts. As a result, RCTs may lose the special status they have gained among social science and policy evaluation methods, turning into one good technique among others.

Keywords Field theory · Global fields · Interstitial fields · Illusio · Policy evaluation · Randomized controlled trials

✉ Malte Neuwinger
malte.neuwinger@uni-bielefeld.de

[1] Faculty of Sociology, Bielefeld University, Bielefeld, Germany

Introduction

Few topics in the social sciences are as hotly debated as randomized controlled trials (RCTs). This is not because drug trial-style experimental studies with randomly allocated "treatment" and "control" groups are more interesting or important than, say, poverty, global political turmoil, or the threats posed by new technologies. It is because RCTs hit so close to home. Over the last decade and a half, many economists, political scientists, and sociologists have begun to claim that well-designed RCTs will greatly improve both the theory and the practice of social science (Baldassarri & Abascal, 2017; Banerjee & Duflo, 2009; Humphreys & Weinstein, 2009), giving rise to what some have called a "credibility revolution" (Angrist & Pischke, 2010).

Even beyond academic social science, RCTs are seen as so useful for getting a handle on real-world problems that three of their main proponents, Esther Duflo, Abhijit Banerjee, and Michael Kremer, have recently won a Nobel Prize. Evidence gained from RCTs is advertised as a new way out of the "ideology, ignorance, and inertia" many policy debates appear to be stuck in (Banerjee & Duflo, 2011, p. 16). Some fifty years after social science methodologist Donald T. Campbell first dreamed of an Experimenting Society based on RCTs, a "twenty-first century experimenting society" seems to be emerging (White, 2019). By now, RCTs in social settings are not only a scientific method—they are also a multi-million-dollar business.

Given such high hopes, bold proclamations, and economic potency, it is hardly surprising that a sizable group of social scientists begs to differ. Critics argue that RCTs are no more rigorous than other techniques (Bédécarrats et al., 2020; Deaton & Cartwright, 2018), limited at best when it comes to addressing real-world problems (Berndt, 2015; Pearce & Raman, 2014), and generally ethically worrisome (MacKay, 2018; Teele, 2014). In this view, RCTs rarely generalize to other places, rely on limited and faulty data, marginalize broader political problems, introduce a technocratic focus into social science, and are in danger of violating people's rights. While to proponents more and better RCTs seem like the only way to go, their critics often find it puzzling that the much-maligned "rationalist" model of policymaking keeps cropping up in yet another disguise (Kelly & McGoey, 2018; Oliver, 2022; Picciotto, 2012).

This article takes arguments for and against RCTs seriously, but it engages with them only after taking a more empirical approach to the recent success of RCTs in scientific and applied contexts. Instead of starting from the premise that they revolutionize credibility or claiming that they fall back on a long-debunked chimera, it asks, *What accounts for the proliferation of RCTs in the first place?* Note that, in this conception, the question of the "success of RCTs" is quite independent of the question of whether they have in fact solved the scientific and political problems they purport to solve, or even whether they can do so in principle. It simply acknowledges that RCTs have spread enormously and asks how this could have happened.

Though having gotten little attention in the heat of the present debate, this approach turns the controversy into a theoretically intriguing research problem

of considerable practical relevance. Theoretically, it zooms in on questions of the interconnection of science and politics and the merging and decoupling of social fields. More practically, it helps to make the debate more level-headed and productive. The article addresses praise and critique of RCTs not by explicitly favoring one side, but by showing that the dominance of RCTs is less extreme than it may seem and that many practitioners have begun to accept critiques and adapt accordingly. This makes it possible to argue that RCTs are a fine method among others, with strengths and weaknesses, and that top RCT proponents at leading institutions are coming around to this view. Debate in the name of advancing social science is legitimate, necessary, and welcome. Yet there is no need to villainize RCTs or be afraid of them—just as it is unhelpful to idealize RCTs or be blinded by the current hype.

Explaining the success of RCTs, it has been said, is "like trying to chart the birth of rock and roll. Early influences are many, and every fan has a story" (Angrist & Pischke, 2010, p. 5). One popular and initially intuitive explanation is to point to their inherent scientific superiority. Particularly proponents claim (or at least imply) that RCTs became popular simply because they are better than other techniques (Duflo & Kremer, 2005; Leigh, 2018). However, several scholars have noted that this view is not particularly convincing (Bédécarrats et al., 2019; de Souza Leão & Eyal, 2019; Donovan, 2018). For one thing, if RCTs had spread purely because of their superiority, why were they not popular all along? Considering that there have been several "waves" of RCTs since the statistician Ronald Fisher popularized them in the 1920s (Jamison, 2019), the inherent superiority hypothesis does a bad job of explaining why each of these waves subsided after a few years.

For another, if one wants to claim that RCTs will "revolutionize social policy" in roughly the same way as they did for "medicine in the twentieth century", as Duflo and Kremer (2005, p. 228) allege, shouldn't one take the actual circumstances of this supposed historical precedent much more seriously? After all, the establishment of RCTs as the "gold standard" of drug testing is a textbook example of strong-state regulatory policy, not of everyone magically being swayed by the power of Reason. Randomized double-blind clinical studies only became standard practice in the 1960s after the German pharmaceutical company Grünenthal had managed to poison thousands of unborn babies through its sleeping pill *Contergan*, otherwise known as thalidomide. This incident put pressure on regulatory agencies worldwide to prevent other pharmaceuticals from doing similar things in the future (Carpenter, 2014). The issue here is not whether mandating RCTs was epistemically warranted—it may well have been—but that the "revolution" in medicine was supported through a level of state regulation previously unheard of that made selling drugs without backing them through RCTs simply illegal. In other words, if most of today's medical professionals believe in the value of RCTs, they began believing only after most states of the industrialized world had declared that not doing so was equivalent to advocating the free exchange of poison (Marks, 2000).

As these examples make clear, this article's main concern is not whether RCTs are a good research technique but (much more modestly) whether they could plausibly have spread solely as a result of their alleged superiority over other methods. The historical record suggests that they hardly could. A more plausible story is suggested

by the title of one recent history of social policy RCTs, written by two proponents: *Fighting for Reliable Evidence* (Gueron & Rolston, 2013). As their emphasis on *fighting* suggests, the insight that the recent proliferation of experimental methods came about through a process mostly unrelated to science proper can be shared by critics and proponents, even while legitimate disagreement about the adequacy of these methods persists. Whether you think that RCTs are unnecessary and unethical or revolutionary and righteous, the question arises: How exactly did the process of RCT popularization unfold?

The argument of this article is developed in close conjunction with two recent answers to this question. One is that RCTs have become part of a "scientific business model" through which researchers, often by establishing elite research networks like the Abdul Latif Jameel Poverty Action Lab (J-PAL), have managed to sell their research to a variety of non-scientific customers (Bédécarrats et al., 2019). The other is that favorable institutional conditions—such as shifts in academic economics and development aid—turned RCTs into "hinges" between formerly disconnected fields, durably linking them together by rewarding RCTs in both academic and applied contexts (de Souza Leão & Eyal, 2019). While both perspectives are valuable, one central difference among them is that the former explains the success of RCTs through the scientific and political savvy of strong actors pursuing their interests, while the latter takes a step back to uncover an institutional dynamic that "rewires" the interests of all actors involved. Both studies describe a political process, but one tells a story of power and influence while the other tells a story of unlikely alliances among former strangers.

This article provides empirical evidence that the truth involves some combination of both accounts, often working in parallel. But while accepting several of their arguments and observations—notably the latter's field theory perspective and its description of developments in economics and philanthropy—it also goes significantly beyond them. The main argument remains that during the 1980s and 1990s shifts toward behavioral economics and small-scale project-based development aid did indeed create key institutional conditions suitable to establish RCTs as "hinges" between scientists and practitioners. What is new is that these favorable conditions were never confined to the development sector, instead gaining ground through an additional—more general—shift toward New Public Management. By the early 2000s, RCTs thus functioned as hinges not only in countries of the Global South but also of the Global North. Through a process of intellectual and political cross-fertilization among numerous fields, the early 2010s then saw the crystallization of what I call a "global interstitial field" (Buchholz, 2016; Eyal, 2013; Medvetz, 2012)—a conglomerate of states, international organizations, NGOs, researchers, and philanthropic foundations in favor of RCTs, connected through a relatively stable social arrangement with institutionalized boundaries and internal hierarchies. United in this shared field, proponents became able to run RCTs in an increasing number of policy areas and garner significant political influence.

Most importantly, the RCT success story comes with a catch. The article shows that cooperation among researchers and practitioners is anything but smooth in practice. Because researchers prioritize publishing papers in academic journals, while policy-makers and funders focus on improving real-world programs and achieving

quick policy impact, contradictions and goal conflicts emerge. The hinges between fields do exist, but they are much weaker than often assumed and somewhat "squeaky". Because researchers and practitioners originate from diverse fields, they also partly "inherit" the *illusio*—the central incentives or stakes—operating in these fields (Bourdieu & Wacquant, 1992, pp. 98–99). This leads to coordination problems among key actors. In this sense, the success of RCT proponents' fight against ideology, ignorance, and inertia depends on their ability to manage the consonances, compromises, and contradictions of the global interstitial field in which they find themselves.

In the sections that follow, the article first reviews recent explanations of the rise of RCTs since the early 2000s. It argues that the two main weaknesses of this branch of research are that they mostly focus on experiments in poor countries and attribute an implausible amount of agency to a small number of "model cases" like J-PAL (Krause, 2021). Therefore, they overlook developments in wealthier countries and downplay the number and diversity of RCT supporters. Recent accounts that interpret experimental methods as a scientific business model or as hinges between fields partly ameliorate these oversights, but also reproduce them in certain respects. After analyzing the crystallization of the global interstitial field of RCT support and the goal conflicts it produces, the article concludes with two tentative predictions. First, support for RCTs will probably differentiate according to the scientific and applied fault lines already perceivable today. Second, and partly as a result, RCTs may lose the special status they have gained among social science and policy evaluation methods, turning them into one good method among others. This does not necessarily mean that the general influence of RCTs will subside, but that their production may become organized much like drug testing or market research are organized today.

Recent explanations of the rise of RCTs

Because debates among social scientists have focused largely on the scientific, political, and ethical pros and cons of RCTs, they have devoted less attention to the empirical question of why RCTs have been spreading in the first place. Implicitly, many probably assume that the latter question is merely a "special case" of the former, in the sense that high popularity is a result of compelling arguments in favor of RCTs. But as I have argued, the popularity of RCTs is at best loosely coupled with arguments speaking in their favor. This section reviews the explanations and empirical studies currently available that acknowledge this point. It argues that most of them share two main starting points, which lead to two main drawbacks.

First, most researchers assume that the current success of RCTs is rooted in applications in the Global South, leading them to neglect experimental evaluations in developed industrialized countries. Second, most researchers treat small research networks like J-PAL as "model cases" that stand in for the proliferation of RCTs in general. Explicitly or implicitly, this leads to the attribution of an implausible degree of agency to a relatively small number of actors, what I call a "baseline individualism" of current research. To some extent, these tendencies even pertain to two of the most inspired contributions to the discussion, namely the claim that RCTs form

part of a new "scientific business model" (Bédécarrats et al., 2019) and that they have created "hinges" between formerly separate social fields (de Souza Leão & Eyal, 2019). This analysis suggests that explanations of the success of RCTs can be improved by considering a larger number of supporters, paying attention to RCTs in Northern and Southern contexts, and further clarifying the broader institutional conditions that made this support possible.

"Model cases" in development research

Whether supportive or critical, many social scientists maintain that the current "wave" of RCTs originated in the new development economics of the early 2000s (e.g. Donovan, 2018; Fejerskov, 2022; Leigh, 2018). One touchstone for this impression is Banerjee and Duflo's influential book *Poor Economics* (2011). In strong rhetoric, the authors argue that the economics of development are better conceived as the economics of poverty—and that most economists in this sub-discipline have done their job poorly. Positioning themselves between supporters and critics of foreign aid, Banerjee and Duflo argue that better science and politics can only be achieved through more and better evidence. And "better evidence", they make clear, usually requires RCTs (Labrousse, 2020). Amplified by prizes and enthusiastic media coverage, their story has been subject to a classic Matthew Effect: a few superstars get the credit for the work of a large community (Merton, 1968). Supposedly, the main push for RCTs came out of the small field of development economics, triggered by an even smaller sub-group of elite innovators.

Of course, the story is not entirely baseless. Household examples are analyses of cash transfer and micro-credit schemes in developing countries (Banerjee et al., 2015), with the Mexican *Progresa* study as an early highlight (Tollefson, 2015). Another major success story used to be RCTs on programs in which African children were "dewormed" off intestinal parasites, though by now the so-called "worm wars" have turned deworming into a more controversial issue (Allen & Parker, 2016). And while the "reproducibility crisis" in the social sciences has raised some further issues (Czibor et al., 2019), concerns about whether experimental results are replicable in other contexts do little to alter the perception that the center and initial trigger for conducting more RCTs is to be found in development economics. Even when critics talk about RCTs as the emergence of an unethical "global lab", their critique is premised on the claim that researchers from the Global North experiment on subjects in the Global South (Fejerskov, 2022).

As a consequence, scholars have rarely connected RCTs in developing countries with their counterparts in wealthy industrialized ones—or vice versa. They rarely discuss that RCTs have been part of American labor market policy since the 1960s and that a veritable research industry used to conduct experimental trials on health insurance, tax schemes, and housing (Berman, 2022; Breslau, 1998; Gueron & Rolston, 2013). Nor have they considered how aspirations to reinvigorate these efforts began to form in the governments of Northern states at about the time as development economists started to build their research networks, particularly in the United States and the United Kingdom. As I will expand on below, the US government's idea to use RCTs for

playing "Moneyball for government" (Nussle & Orszag, 2015) and take a behavioral approach to public policy may be considered as at least as important for shaping the RCT agenda as the concerns of academic economists. Though scholars of public policy are well aware of these trends (Haskins & Margolis, 2015; Jones & Whitehead, 2018; Pearce & Raman, 2014), they have rarely related their work to the question of the initial success of RCTs. If scholars do note the connection, they tend to tacitly agree that experimental methods first emerged in international development (e.g. Jones & Whitehead, 2018).

Connecting to the assumption that the present wave of RCTs first emerged in the Global South, the second assumption broadly shared among social scientists is that the current wave of RCTs is led by a small number of newly established research organizations (e.g. Fejerskov, 2022; Jatteau, 2018; Karlan, 2011). The first on everyone's list is the Abdul Latif Jameel Poverty Action Lab (J-PAL), occasionally accompanied by Innovations for Poverty Action (IPA) and the World Bank's Development Impact Evaluation office (DIME). The Bill & Melinda Gates Foundation, the William and Flora Hewlett Foundation, and the UK Department for International Development (DFID, now FCDO) also sometimes receive an honorable mention, largely because they provide the necessary funding (Donovan, 2018). But this is about the most detailed things get.

In this sense, recent research has turned a few select organizations into "model cases" that *stand in* for a much larger epistemic target (Krause, 2021), namely the success of RCTs *in general.* Especially J-PAL and its Nobel Prize-winning founders have become the starting point for explaining the phenomenon of experimental trials, and they are the privileged research object for interested scholars. Paradoxically, insofar as social scientists have been able to say anything about the rise of RCTs, this general depiction is achieved by looking at a particularly narrow set of examples.

Model cases are useful to focus scholarly attention on particular research objects and sites, but they also trigger analytical problems. While support for RCTs, according to one scholar, consists of "a dizzying array of initiatives and organizations" (Donovan, 2018, p. 30), in actual research practice the focus on model cases leads to extreme selectivity regarding actors considered truly relevant. For instance, another scholar presents "the elitism of the J-PAL and the tightened network of randomists as an explanation for the success of RCT" (Jatteau, 2018, p. 115), hence leaving all other initiatives of the "dizzying array" out of the picture. The main downside of turning a few heroic innovators into model cases is that it leads to a strong "baseline individualism": while no one seriously claims that focusing on J-PAL tells the full story about the spread of RCTs, the fact that J-PAL is the only actor that has been seriously researched makes scholars fall back on the familiar one-dimensional story. Before correcting these oversights, it makes sense to investigate two recent explanations of the rise of experimental methods in some more detail.

A new "scientific business model" and the emergence of "hinges" between fields

The weaknesses of recent explanations of the success of RCTs—neglect of its broader international scope and a baseline individualism—are also present in two of

the most insightful articles on the topic: Bédécarrats, Guérin, and Roubaud's (2019) claim that RCTs form part of a new "scientific business model" and de Souza Leão and Eyal's (2019) proposal that experimental methods establish "hinges" between the fields of academic economics and practical development work. Acknowledging this is useful not only to show that the weaknesses are real but also to suggest how they may be overcome. While especially the former explanation remains strongly individualistic and both neglect experimentation in the Global North, I argue that they are the most plausible approaches we currently have. Extending and relating them to one another thus provides the basis for the main argument of this article.

Rooted in political economy, the main strength of Bédécarrats and colleagues' argument is to connect a scientific trend like experimental methods with political and economic interests. It argues that leading social experimenters, led by future Nobelists Duflo, Banerjee, and Kremer, "have generated an entirely new scientific business model, which has in turn driven the emergence of a truly global industry" (Bédécarrats et al., 2019, p. 752). Young researchers "from the inner sanctum of the top universities" (ibid.) have managed to combine the "mutually reinforcing qualities of academic excellence (scientific credibility), public appeal (media visibility and public credibility), donor appeal (solvent demand), massive investment in training (skilled supply) and a high-performance business model (financial profitability)" (Bédécarrats et al., 2019, p. 752).

To make the business model function, researchers have set up NGOs like J-PAL and IPA, which can receive funds from a variety of sources: support comes not only from public research funding but also from philanthropic foundations and businesses. As a consequence, researchers and their NGOs "have created an oligopoly on the flourishing RCT market", including a large field infrastructure necessary for conducting RCTs (Bédécarrats et al., 2019, p. 753). Overall, Bédécarrats and colleagues construct a straightforward model involving a group of powerful actors who have mobilized their economic, cultural, and social resources to gain enormous scientific influence and operational capacity. The result of these efforts is a scientific business model that profits from conducting RCTs and supporting their perceived superiority. What makes this explanation somewhat problematic, though, is that it largely rests on the assumed influence of a small group of researchers and organizations assumed to be all-powerful (again, a case of baseline individualism resulting from reliance on model cases) and that the "global industry" they have supposedly created does not include RCTs in the Global North.

Rooted in political sociology, de Souza Leão and Eyal's approach focuses less on uniquely powerful individuals and more on the broader institutional infrastructure that is necessary to support them. Substituting Bédécarrats and colleagues' leading analytical concepts, "market" and "interest", in favor of "fields" and "hinges"—in the sense of Bourdieu (1985) and Abbott (2005)—enables the sociologists to reconcile the research industry's expansion with its broader social and political environment. As they put it,

> the contemporary success of RCTs is better understood as a product of historical and institutional processes that have changed the political and scientific context in which RCTs are implemented, rather than as evidence of their "gold

> standard" quality. By jointly mobilizing the concepts of "hinge" and "homology between fields", we show how the fragmentation of the development aid field and changes in the economics profession made RCTs answerable to new audiences and allowed *randomistas* greater leeway to bypass the political resistance to randomization (de Souza Leão & Eyal, 2019, p. 412, emphasis in original).

Phrased differently, the popularity of RCTs clearly cannot be explained purely through their alleged scientific superiority, but pointing to the strategic efforts of heroic innovators is not sufficient either. This is so because these heroic efforts relied on historical and institutional conditions that needed to be established first. The observation that organizations like J-PAL were established by the beginning of the 2000s to promote RCTs is correct, but it merely pushes back the deeper question of why these organizations were founded at this point in the first place.[1]

The key institutional condition that had to be established (and that is missing in the story of the scientific business model), de Souza Leão and Eyal argue, was that the previously relatively independent fields of practice-oriented development aid and theory-oriented academic economics became *durably linked*. This became possible through "homologous transformations of development aid and economics" that took place during the 1980s and 1990s (de Souza Leão & Eyal, 2019, pp. 401–408). In academic economics, disciplinary shifts toward behavioral approaches and an emphasis on causal attribution made experimental studies about people's actual economic decision-making an attractive research topic (Sent, 2004). Because such experimental studies are easiest to set up in development contexts, development economists had an incentive to team up with development NGOs (Berndt, 2015).

In addition, the gradual dissolution of the Washington consensus—the market liberalization agenda that had dominated development policy for the past two decades (Rodrik, 2006)—fragmented the field of development aid, creating space for new actors, especially philanthropic foundations and NGOs. Together, these institutional transformations led to a situation in which academic researchers for the first time faced career incentives to do messy and time-consuming practical experimental research, while the influence of philanthropic foundations encouraged development NGOs to focus on "clear goals" and "measurable results" (de Souza Leão & Eyal, 2019, p. 405). Taken together, these institutional developments turned RCTs into a hinge: they had come to connect two previously disconnected fields because *conducting experimental trials now provided rewards in scientific and applied contexts*.

Overall, de Souza Leão and Eyal's argument that the incentives of leading supporters of RCTs are embedded in broader social fields and enabled through links between them is a unique way of overcoming the baseline individualism present in the literature. Unfortunately, even this broader analysis focuses exclusively on developing

[1] De Souza Leão and Eyal do not directly refer to Bédécarrats and colleagues but instead criticize an excellent article by Donovan (2018). While Donovan's argument is based in yet another analytical school, namely science and technology studies, Bédécarrats and colleagues' article is structurally similar with respect to the strong emphasis it puts on the intentionality of a small number of actors.

countries. The remainder of this article thus attempts to retain the main strengths of the explanations just summarized and extends them where necessary.

Methods and data

This article is based on historical reconstruction, an analysis of the global interstitial field of RCT support as a social network, and interviews with researchers and practitioners at leading RCT-supporting organizations. Its historical analysis brings together literature from a variety of fields of study, occasionally extended through reports written by RCT supporters, most of which are freely available online. To accompany this narrative, I compiled a dataset of formally organized supporters of RCTs, which are connected through ties of collaboration and economic dependency. Using known key supporters as starting points—J-PAL, IPA, DIME, the Behavioural Insights Team (BIT) in the UK, and others—I traced the connections between the many "partners" and "funders" organizations advertise in annual reports and on their websites. Yielding 344 organizations, this approach is no doubt limited, perhaps most obviously through self-reporting (which strongly favors NGOs). Nevertheless, it provides a basic overview of the global field of RCT support and the interconnections among its main actors. Finally, the article draws on twenty open-ended interviews with employees of key RCT supporters, including J-PAL, DIME, and BIT, enriched through conversations with several economists and government officials who have spoken in favor of RCTs. All interviews were conducted via video call in 2022 and 2023, lasting between 45 minutes and two hours.

Emergence and crystallization of a new global field

In the very early 2000s, RCTs were obscure. In the United States, "demonstrations" of welfare programs had kept the idea of social experimentation alive during the 1990s. Still, RCTs remained at the fringes of welfare policy and were largely unknown in most areas of government (Gueron, 2017; Harvey et al., 2000). In development policy, as Esther Duflo recalls, experimental trials were the kind of "project that crazy people do in the back yard" and largely the opposite of "something that is institutional and serious" (quoted in Parker, 2010). How did RCTs go from obscurity to seriousness in such a short amount of time?

Attempting to synthesize the arguments of recent research, this section argues that the RCT success story must be understood as an interaction of favorable institutional conditions with the agenda of savvy actors who managed to develop the production and dissemination of RCTs into a scientific business model. During the 1980s and 1990s, New Public Management emerged as a promising concept to reform the public sector through the adoption of business methods. By the early 2000s, this thinking led several Northern governments to require experimental methods as an instrument of "performance measurement", often

establishing RCTs as hinges between academic and applied contexts. Through cross-fertilization among states, NGOs, research institutes, international organizations, and philanthropic foundations—operating in the Global North and the Global South—by the 2010s support for RCTs then gradually established a larger followership, eventually crystallizing into a "global interstitial field" of its own.

Institutional conditions: Hinges in the Global North

What are the institutional conditions that enable RCTs to function as hinges between previously disconnected actors and fields? As described previously, de Souza Leão and Eyal focus on what they call the "homologous transformations of development aid and economics": the dissolution of the Washington consensus opened a door for philanthropic foundations and NGOs to play a larger role in development policy, and the rise of behavioral economics created incentives for economists to use development projects as an opportunity to do empirical work. RCTs turned into a hinge because they provided rewards in both academic and applied contexts. So far, the story is quite convincing. Unfortunately, it focuses exclusively on the Global South and neglects parallel developments in public policy thinking in wealthy industrialized countries. As we will see, behavioral economics created similar incentives for researchers in the Global North, which were matched by homologous transformations in the field of public policy. What this amounts to, I argue, is that the hinges RCTs have created and the institutional conditions their success relies on are more encompassing than acknowledged in previous research.

One factor many researchers and practitioners regard as important for the success of RCT is the rise of New Public Management (NPM), though the exact nature of this influence is rarely specified in detail (Bédécarrats et al., 2019, pp. 750–751; Vedung, 2010, pp. 273–374; White, 2019, p. 2). Mostly a discussion of the 1980 and 1990s, NPM was a loose collection of ideas to reorganize public sector management, reporting, and accounting and bring them "closer to (a particular perception of) business methods" (Dunleavy & Hood, 1994, p. 9). This essentially involved a focus on (1) transparent accounting principles based on "outputs" measured by quantitative performance indicators, (2) decentralization of public services (3), the desire to link employees' incentives to performance, and (4) opening up the provision of public services to competition (Dunleavy & Hood, 1994; Page, 2005).

Because NPM was always accompanied by controversies regarding its politics and practical applicability, by the mid-1990s initial excitement largely gave way to more pragmatic views. Scholars regarded the essence of NPM as "so omnipresent" in Northern anglophone countries "that it hardly amount[ed] to a distinctive reform programme at all any more" (Dunleavy & Hood, 1994, p. 10). NPM was considered a "somewhat dated label" (N. Manning, 2001, p. 197) most people had heard about but which described little beyond a vague pro-market orientation in public administration. Still, by the early 2000s most wealthy European and Anglo-American countries had established New Public Management-oriented reforms—for instance, public service providers would compete against the private sector and adopt a customer rhetoric toward citizens (Page, 2005; Schedler & Proeller, 2002, pp. 165–166).

The spirit of NPM had effects on many areas of administration. In development policy, the adoption of the UN Millennium Development Goals (MDGs) in 2000 and the Paris Declaration on Aid Effectiveness in 2005 emerged under its influence (Pamies-Sumner, 2015, pp. 9–10; White, 2019). More importantly for the present context, public administrators' desire to measure performance also converged with academic work, turning RCTs into a hinge with Northern public policy. One example is the establishment of the US Institute of Education Sciences (IES) in 2002, a new scientific agency with independent authority for knowledge reporting and research funding in the education sector. IES declared RCTs to be the most rigorous research design for evaluating national education programs and aligned the funding requirements of education research with those provided by the Department of Health, meaning that "federal funding for education evaluation shifted almost entirely to randomized trials" (Orr, 2018, p. 55).

Within a year of existence, IES and its newly founded What Works Clearinghouse—an online library functioning as an intellectual grounding for the newly established movement of "evidence-based education"—thus changed the criteria for "rigorous" research, adapted relevant funding mechanisms, and established a new way of disseminating insights about "what works" in education (Haskins & Margolis, 2015, pp. 7–10; Hedges & Schauer, 2018, p. 272; Whitehurst, 2018). As one evaluation expert remarks, the presidency of George W. Bush had brought about "a perfect storm" of RCT-based evaluation, linking executive branches of the US government with education research (Donaldson et al., 2010, p. 33). Especially the United Kingdom and the Scandinavian countries followed the US example a few years later (Dawson et al., 2018; Pontoppidan et al., 2018).

Another well-known example for the convergence of NPM thinking and scientific reasoning in public policy is the rise of "libertarian paternalism" (Thaler & Sunstein, 2003), later rebranded as "behavioral public policy" or simply "nudge" (Halpern, 2015), which also emerged during the early 2000s. The stated goal of this movement was to find an appropriate balance between the "libertarian" principle of people's freedom of choice and the "paternalist" principle of maximizing their welfare through state intervention. Normatively, future economics Nobelist Richard Thaler and lawyer Cass Sunstein argued that because every social arrangement requires implicit choices by an authority—even the order in which food is presented in a cafeteria—carefully "nudging" people in a direction that maximizes their welfare is an acceptable combination of libertarian and paternalist principles. Empirically, drawing on behavioral economics, they argued that because people are systematically biased in their decision making they often fail to do what is best for them. Therefore, mildly paternalist interventions were not only inevitable but *necessary*: to protect people from making bad decisions, default settings, reminders, particular framing of information, and other techniques were required to push people toward beneficial behavior (Thaler & Sunstein, 2003, pp. 175–176).

Libertarian paternalism quickly took off. In 2009, Sunstein, a former University of Chicago Law School colleague of Barack Obama, became Administrator of the White House Office of Information and Regulatory Affairs (OIRA), using his tenure to popularize nudging in the US government (Halpern, 2015). In a TED talk the following year, future British Prime Minister David Cameron (2010)

outlined his political vision of a "next age of government" in the aftermath of the 2008 financial crisis. Asking, "How do we make things better without spending more money?", Cameron opted for lower-cost government based on insights from behavioral economics. One key result of these developments was the founding of the British Behavioural Insights Team (BIT) in 2010, known as "one of the most significant fusions of the behavioural sciences and government witnessed anywhere in the world" (Whitehead et al., 2018, p. 41). By 2013, 135 countries had adopted some form of behavioral public policy, and 51 had centralized these operations in a more structured, government-led form (Whitehead et al., 2014). Soon, the World Bank (2015), the United Nations (2016), the European Union (2016), and the OECD (2017) all began to promote nudging policies.

The notable aspect of these developments is not only the prevalence of NPM reasoning but also that the scientific theories of behavioral economics and the political agenda of libertarian paternalism were again held together through the hinge of RCTs. Though there are exceptions, the style of libertarian paternalism that got the most political traction—the one popularized by BIT—usually tests its interventions through experimental methods (Ball & Head, 2021; Lee & Ma, 2020). To frame libertarian paternalism as pragmatic and common-sensical, advocates needed a kind of evidence that was able to empirically demonstrate both the *effectiveness* and the *cost-effectiveness* of nudges (Einfeld, 2019; John, 2018, pp. 5–6).

This was necessary because nudging policies frequently drew the criticism of a manipulative "psychological state" (Whitehead et al., 2018, p. 26), and even practitioners themselves acknowledged that available academic studies supporting behavioral approaches were less solid than often assumed (Jones & Whitehead, 2018, p. 318). As an intuitive tool that seemed to demonstrate beyond reasonable doubt that behavioral interventions are both self-evident and evidence-based, RCTs could be used to defend nudging against criticism and to establish political trust (Whitehead et al., 2018, pp. 25–26). In addition, RCTs also provided libertarian paternalists with a way to confidently claim that a certain policy intervention saved a particular amount of money, nicely connecting with NPM concerns. By extrapolating the effect sizes measured in one RCT to the whole population, a 2012 Behavioural Insights Team publication even claimed that the team had "achieved savings of around 22 times the cost of the team and identified specific interventions which will save at least £300 m over the next 5 years" (BIT, 2012, p. 1). Again, the RCT hinge moved libertarian paternalism away from slightly manipulative behavioral research and toward pragmatic "smart" governance.

Overall, these examples show that RCTs began to function as hinges between political and scientific endeavors in the Global North at about the same time as they did in the Global South. In all cases, a symbiotic relationship between science and politics emerged. RCTs were crucial for the success of NPM, and NPM supported the success of RCTs. Further institutional conditions could be considered. For instance, during the 1990s the movement of evidence-based medicine (NPM) first made a strong case for basing therapeutic decisions on RCTs. By the early 2000s, NPM became explicitly recognized as a model for evidence-based public policy (Daly, 2005; Davies et al., 2000). In any case, the examples discussed here

are enough to demonstrate that NPM was an important institutional condition for the rise of RCTs in countries of the Global North, and they also demonstrate that RCTs were a crucial tool for merging the homologous transformations toward managerialism and behavioral science into a coherent movement. I now ask to what extent this movement has turned into a global field of its own.

Crystallization of a global interstitial field: The cross-fertilization of RCT support in Northern and Southern countries

Social fields are not just any kind of social space. They are the parts of social space that have developed institutionalized boundaries on the outside and structural hierarchies on the inside. Actors can be part of a field or not, and field members can have more or less power over the field's central concerns and future trajectory (Wacquant, 2019; Wacquant & Akçaoğlu, 2017, pp. 61–64). By this definition, has support for RCTs turned into a field? Are all supporters part of the same field? Is the scope of this field global? My claim is that the answer to all three questions is yes. As I argue, the institutional conditions described in the previous section first led to developments in Northern and Southern countries that were only loosely related. But by 2010 these trends merged, as an increasing number of organizations supporting RCTs developed collaborations and economic ties. Ministries of national governments, NGOs, international organizations, research institutes, and philanthropic foundations came to crystallize into a "global interstitial field" (Buchholz, 2016; Eyal, 2013), linking and overlapping with more established fields of politics, academia, and business. Their "in-between" position releases members of interstitial fields—such as think tanks (Medvetz, 2012)—from some of the constraints academics or politicians face. As will become clear later, the liminal position RCT supporters have created for themselves is simultaneously their greatest strength and their greatest weakness.

The favorable institutional conditions that had become established by the early 2000s affected many existing social actors—and also created new ones. By that time, however, the only visible hinges RCTs had established between academic and political fields were very localized, largely remaining confined to education and labor market research in the United States (Angrist & Pischke, 2010; Breslau, 1998). Indeed, one economist interviewed for this study ventures that if researchers interested in RCTs "had not found the international niche, I think they would have been doing it in the domestic space" (interview economist 2). Here my analysis shifts from de Souza Leão and Eyal's focus on institutional conditions to a greater focus on the agency of savvy actors, as argued by Bédécarrats and colleagues. Initial excitement about RCTs was possible only under the right conditions, but as these conditions had emerged key actors took the lead to establish RCTs as a "scientific business model". This is not to imply that, at this point, institutional factors stopped being important. It merely stresses that movement leaders like J-PAL gained influence only after the conditions were right. Figure 1 depicts a rough trajectory of the field's emergence and growth over time. As its leaders established an increasing

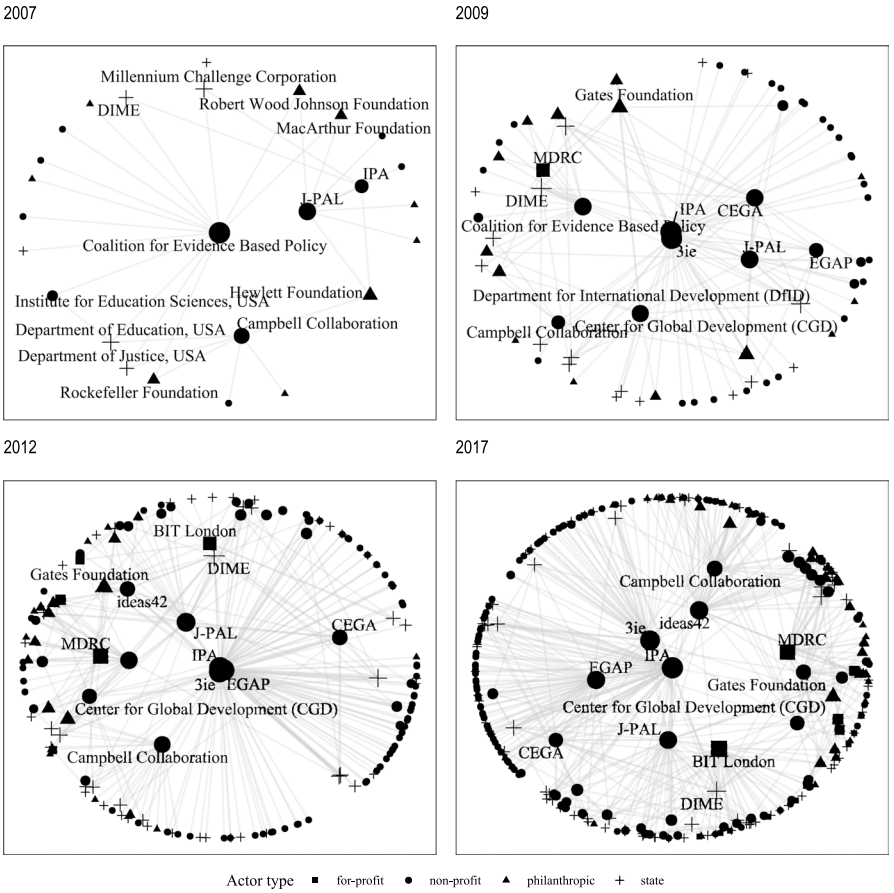

Fig. 1 Four snapshots of the developing global interstitial field of RCT support. Organizations are connected if they collaborated on experimental trials *or* financed each other for purposes of RCTs, each over several years. Particularly collaborative and well-financed organizations appear closer to the center while the rest mark the periphery

number of collaborations and financial ties, they were joined by a growing number of supporters. Gradually, a social structure emerged. Some actors were in, some were out. Some set the agenda, others followed.

The first key institutional move toward RCTs came from the US Office of Management and Budget (OMB), the executive office responsible for making sure that government agencies' activities comply with the president's political line. In 2001, OMB introduced the Program Assessment Rating Tool (PART), a procedure intended to link budget decisions to "program performance" by grading programs from "effective" to "ineffective" (Haskins & Baron, 2011, p. 8; Moynihan, 2013). In part because of effective lobbying from the Coalition for Evidence-Based Policy, a 2004 document titled "What Constitutes Strong Evidence of a Program's Effectiveness?" (OMB, 2004) clarified that OMB's yardstick for "performance" was

RCTs. OMB thus became the unlikely "quarterback of evidence-based policy making" inside the Bush administration (Stack, 2018, p. 112). Lobbying OMB was a big deal because it meant politically linking funding decisions to the use of a particular research method. As another interviewee put it, "When money was at stake, suddenly everybody learned what a randomized trial was, in the nonprofit community and elsewhere" (Interview CEBP). These developments were further strengthened with the inauguration of the Obama administration in 2008, culminating in what economist and former OMB president Peter Orszag and his predecessor Jim Nussle call "Moneyball for government" (Nussle & Orszag, 2015). Obama's stimulus packages, made available against the fallout of the global financial crisis, allowed OMB to increase its evaluation capacity, provide technical assistance to ever more branches of government, and in many cases tie funding decisions to RCTs (Haskins & Margolis, 2015; Stack, 2018, pp. 117–119).

Over the same timeframe, support for RCTs also became increasingly strong *outside* the US government. In 2001, Peter Rossi, Fred Mosteller, and Robert Boruch established the Campbell Collaboration. In 2002, Dean Karlan founded Innovations for Poverty Action (IPA), and in 2003 Esther Duflo, Abhijit Banerjee, and Sendhil Mullainathan started the Abdul Latif Jameel Poverty Action Lab (J-PAL). The year 2005 saw the establishment of the World Bank's Development Impact Evaluation unit (DIME), followed by the International Initiative for Impact Evaluation (3ie) in 2008. These NGOs, international organizations, and research networks are generally regarded as key RCT supporters, far more consequential than the US government (Bédécarrats et al., 2019; de Souza Leão & Eyal, 2019; Donovan, 2018). Yet the reality is more complex. The early 2000s saw a *cross-fertilization* among government circles, development researchers, philanthropists, and NGOs. The US government's domestic concerns to evaluate performance and its experience with RCTs provided initial fertile ground for the establishment of NGOs like J-PAL—but by the late 2000s, as the success of the latter players became evident, excitement "looped back" toward the domestic policy space and larger government-backed organizations more generally.

Another example of this cross-fertilization among domestic politics, academic economics discourse, and development work is the establishment of the Millennium Challenge Corporation (MCC). Set up in 2004, MCC was designed as a crucial NPM-inspired reform of US development policy, namely to directly link foreign aid to developing countries' willingness to enact market-based and democratic reforms (Hook, 2008). Yet the impetus for MCC to focus on RCTs emerged through a highly influential 2006 report by the Center for Global Development, titled *When Will We Ever Learn?* (Sturdy et al., 2014, p. 438). Written by leading supporters of RCTs—among others, Esther Duflo, World Bank Chief Economist François Bourguignon, and Gates Foundation Chief Economist and future USAID Administrator Raj Shah—the report argued that the key problem of development policy was an "evaluation gap" that made it impossible to assess to what extent a policy was having the causal "impact" it aimed for (CGD, 2006).

At a time when J-PAL and IPA were still in their infancy, RCT supporters focused on persuading state-backed development actors of the value of RCTs. But as the newly founded research networks gradually built up their intellectual reputation, their influence became more direct. As Esther Duflo describes

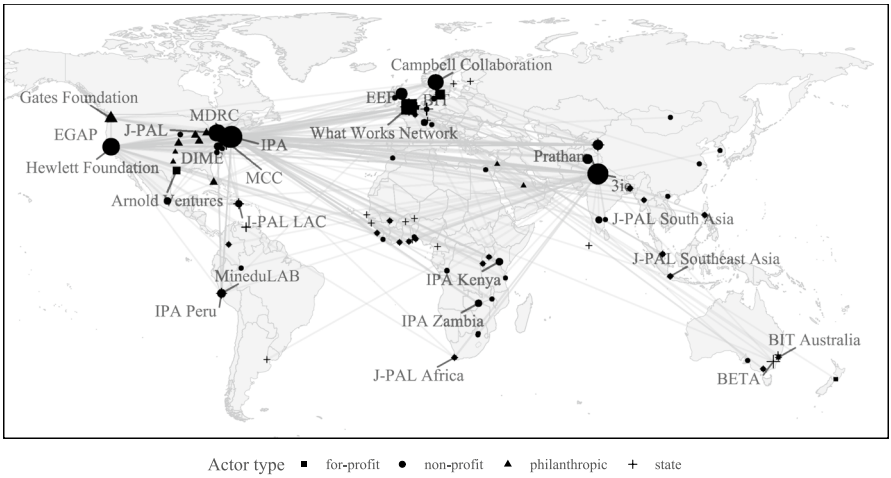

Fig. 2 The global field of RCT support, as of 2021, arranged on a world map. Organizations are connected if they collaborated on RCTs *or* financed each other for the purposes of conducting them, each over several years. Organizations' geographical locations are based on their head office. *Note*: Because so many organizations are based in certain global centers (particularly London and New York, but also others), there is significant over-plotting in these areas

it, RCTs "suddenly became a way to do business […], with academics starting their own projects or starting to participate in large projects" (Duflo in Gueron & Rolston, 2013, p. 466). Only when this "business" showed potential, during the second half of the 2000s, did international organizations like the World Bank manage to acquire large-scale funding for RCTs (DIME, 2010, p. 50). This let DIME grow from "maybe half a dozen staff in total" in 2010 to about 300 today (interview DIME).

The early 2010s are the time when the cross-fertilization among RCT supporters eventually crystallized into a relatively stable global field of its own, durably linking diverse actors around the world under the leadership of a set of NGOs (J-PAL and IPA), international organizations (World Bank), and key philanthropic financiers (particularly the Gates and Hewlett Foundations and, later, Arnold Ventures). Various anglophone countries began to establish organizations doing RCTs, sometimes relying explicitly on the role model of US public policy (AUE & Nesta, 2011; Ball & Head, 2021; Pearce & Raman, 2014). IPA and J-PAL started their first country offices in Africa and Asia, establishing the unequal North–South research relations today criticized as a "global lab" (Fejerskov, 2022). Hundreds of smaller players followed their lead. Recent highlights of the global field's expansion are the World Food Programme's 2019 Impact Evaluation Strategy (World Food Programme, 2019) and the 2022 nomination of IPA founder Dean Karlan as Chief Economist of USAID. By the late 2010s, RCTs had durably connected the "Moneyball for government" project of the US with the "scientific business model" of academic economics. A confluence of efforts, led by leaders in the US and Europe and increasingly finding followers all over the world, had turned RCTs into a global success story (Fig. 2).

How strong are the hinges? Compromises and contradictions in the global field

The argument so far has traced the success of RCTs to a heterogeneous group of academically successful, affluent, applied, and media-savvy supporters who, in conjunction with favorable institutional conditions, managed to establish durable links among previously unconnected fields. This argument has attempted to synthesize previous research on the rise of RCTs and extend it where necessary, particularly emphasizing the cross-fertilization of RCT support in Northern and Southern countries and the dual rewards RCTs began to promise in academic and applied contexts.

However, skeptical readers may have wondered whether this story might not be a bit too neat. Is it really plausible that a global field, premised on support for RCTs, could emerge without internal conflict, coordination problems among key players, and the kind of bad luck most people experience once in a while? This concern is more than justified. Perhaps the greatest weakness of current research is that it tends to present the rise of RCTs as an unstoppable avalanche. While some scholars have moved away from the assumption that RCTs have spread because of their internal superiority, their exclusive focus on the movement's expansion is in danger of establishing another tautological story in which RCTs necessarily come out on top.

Even many critics, who over the past decade have elaborated important epistemic, political, and ethical problems of the RCT movement (Bédécarrats et al., 2020; Deaton & Cartwright, 2018; Teele, 2014), seem less interested in discussing what their criticism has achieved than in puzzling over why RCTs keep spreading despite having solved few of the political problems they meant to solve (Devaux-Spatarakis; Neuwinger, 2023). Yet the reality is that the accumulation of criticism, and even more so the practical difficulty of conducting RCTs and attempting to change real-world decision-making, is putting supporters under increasing pressure (Ball & Head, 2021; Williams, 2023). As we will see in the following section and the conclusion, this pressure leads them to gradually adapt their position and accept common critiques. Naturally, this acceptance and adaptation occurs slowly, grudgingly, and remains somewhat under the surface. Yet it is happening, and critiques of RCTs have played no small role in the recent shift.

My main argument, however, is that the hinges RCTs have established among researchers and practitioners are weaker and less coherent than often thought. Instead of linking fields "seamlessly", as de Souza Leão and Eyal (2019, p. 405) argue, the hinges making up the interstitial field are often fragile and, in some cases, contradictory. Academics do face incentives to team up with practitioners and conduct experimental trials—but their desire for innovation and novelty does not quite fit with the more mundane demands of regular program evaluation. Governments do want to demonstrate that their domestic policies and development aid are backed by "rigorous" evidence—but this usually involves long-term commitments to real-world policies and programs, clashing with academics' desire to test exciting new approaches. And funders, especially philanthropic donors, do

have affinities with testing policies like companies test new products—but they have little patience with the frequent "null results" experimental evaluations tend to produce. While hinges between previously disconnected fields make the global interstitial field possible, they also create goal conflicts. This might be a more general consequence of the functioning of interstitial fields (Eyal, 2013; Liu, 2021).

Tensions between researchers and practitioners: Interesting publications vs. addressing real-world problems

The theory of a functioning hinge is ingenious and attractive. As de Souza Leão and Eyal (2019, pp. 401–402) argue, because RCTs have become valuable to economists, government practitioners, and philanthropists, all of these diverse social actors have an incentive to contribute to the movement. By the early 2000s, doing RCTs started to provide dual rewards for academics (in the form of publications) as well as for practitioners and funders (in the form of "rigorous evidence"). In the language of field theory, RCTs function as hinges because they align the *illusio* of distinct fields—the main incentives or stakes to which actors are exposed (Bourdieu & Wacquant, 1992, pp. 98–99)—enabling the emergence of an interstitial field in which everyone can cooperate effectively.[2] Hinges therefore eliminate the problem of converting institutionalized resources—or "capital"—relevant in one field to resources relevant in another field.

I want to stress, however, that this argument cuts both ways. Because RCT supporters originate from the fields of science, politics, and business, they also "inherit" some of the central incentives and stakes relevant inside these fields. This is because the global interstitial field's relative level of autonomy—the extent to which it can develop field-specific logics, practices, and modes of relevance of its own (Buchholz, 2016, pp. 36–40; Krause, 2018, pp. 8–11)—remains limited. While I have argued that the community of RCT supporters has indeed crystallized into a field of its own, more autonomous and settled fields keep projecting their logic and criteria of relevance on the interstitial field. In this situation, the incentives of the diverse field members become imperfectly aligned—they are torn between different field-specific logics and criteria.[3] The hinge between researchers and political practitioners provides a first example.

Over the course of the 2000s, doing RCTs had become a way for academic researchers to get published in prestigious journals. Pulling off an experimental study had turned into a central marker of skill and "rigor" (Bédécarrats et al.,

[2] Note that *illusio* has nothing to do with illusions or false beliefs. Instead, it derives from the latin *ludus* (english "game") and describes a state of mind in which people are really involved in the "social game" they are playing and accept its incentives and stakes as important (Bourdieu & Wacquant, 1992, pp. 98–99). If a researcher leaves the scientific field and starts working in a real job, publishing papers and getting quoted by fellow academics does not become unreal—it just becomes unimportant.

[3] This argument is somewhat akin to investigations of clashing "institutional logics" (Thornton & Ocasio, 2008), though I would argue that the field concept is clearer about outside boundaries and inside stratifications.

2019, p. 754; Gërxhani & Miller, 2022). But because the incentives of academics rarely align with those of practitioners, conducting RCTs in applied contexts is often difficult. As one World Bank researcher explains, for DIME evaluators (and even more for researchers with the Bank's Development Research Group) "the metric you're evaluated on is, like, 'how many papers do you publish?', which pushes towards that academic side, and pushes away from being really responsive to the concerns and questions from the operational team" (Interview DIME). Similarly, another economist laments that her colleagues tend to "create the questions, instead of thinking, you know, 'we should be there in service of the problems and questions that the practitioners have"' (Interview economist 2). And an evaluator with the German KfW Development Bank remarks,

> *KfW:* Academic work is often rather detached from real practical work, of the things that are really going on in terms of programs. Ideally, this shouldn't be so, but it is. And the reason is of course: The scientific world incentivizes great RCTs, funky designs, new data, precise identifications. And this is easier to get if I [as an academic] do my own experiments. But actual practical work depends on technical solutions that depend on real-world situations. And reality often may not be interesting enough that you can do a great RCT on it, or anything that you can publish in a top journal. But as you know, this is what young researchers need to do.

From this perspective, the global interstitial field and its scientific business model suffer from a clear internal divide. In their everyday work, researchers are primed to focus on novelty in their publications while practitioners must aim for practical improvements of existing policies and programs. In the words of one researcher who works for the philanthropy Arnold Ventures,

> *Arnold Ventures:* I mean, [at research-focused organizations] there's certainly an effort and an attempt to ensure that the questions that are being asked are the questions that implementers or governments would want the answers to. But at the end of the day, the projects that get developed at, like, the IPAs and the J-PALs of the world are driven by academics. I think that's a real difference from what we're trying to do here.

The existence of tensions between the *illusio* of actors operating primarily in scientific rather than applied contexts, and vice versa, leads to one additional insight into the dynamics of imperfectly aligned fields. De Souza Leão and Eyal (2019, pp. 404–405, 398) argue that part of the homologous transformations that enabled the hinges among academics and practitioners to emerge was that current RCTs are based on small nudges rather than large-scale government interventions. As they see it, both groups of actors have converged on a view according to which small changes, tested through small-scale RCTs, may lead to big improvements—an idea that has been called "radical incrementalism" (Halpern & Mason, 2015). But quite to the contrary, interviewed researchers suggest that evaluations of real social programs and small-scale academic "funky designs" are rarely the same (Interview KfW). Because many academic researchers are "on the tenure

clock", hoping to get a university professorship, long-term RCT evaluations on real-world projects are rarely pursued and left to corporate actors with fewer time and funding constraints like the World Bank (Interview DIME). As one expert at the German Institute for Development Evaluation puts it,

> *DEval*: When I think of the Banerjee's of this world, they have their three institutions and three countries they work with. And with those they develop fancy interventions that produce nice publications. But this is not the stuff that USAID or FCDO [i.e. the US and UK aid agencies] need.

These comments suggest that the hinges between academics and practitioners do indeed exist, but they should rather be regarded as the *smallest common denominator* academics and practitioners can agree on. Rather than the result of a perfect confluence of methods, worldviews, and practical needs, small-scale "nudging" RCTs are a compromise necessary to overcome fundamental differences among scientific and applied fields (White, 2014, pp. 21–22). The diverging *illusio* of different fields does not entirely unhinge the linkage provided by RCTs, but the hinge that actually exists is squeaky at best.

In some cases, this "squeakiness" has downright bizarre implications. As a large funder of RCTs, 3ie had agreed with the Mexican government to find qualified academic evaluators for one of its social programs. But as a 3ie employee explains,

> *3ie*: This team bid to evaluate [the program]. And they said, "Well, the only design we can think of is this design. And Esther [Duflo] and Abhijit [Banerjee] already published a paper with this design for programs in India. So we don't want to do that and publish it because that design has been used already. So we're not gonna do it." I'm like, "But you agreed about evaluating this program. We don't care what design you use, just use a valid design". And they said, "No, we're not going to do it, it's not gonna be publishable".

This episode drives home the conundrum of imperfect hinges in interstitial fields. Academics prioritize novelty and originality while governments prioritize real-world improvements and long-term commitment. The hinge turns out to be so fragile that it breaks under the weight of misaligned *illusio*.

Tensions between researchers and funders: Learning what works vs. the requirement of quick success

Having discussed the hinge between academics and political practitioners, it is also worth looking at the hinge between academics and funders. According to recent research, the particular strength of the scientific business model is that RCTs receive funding not only from public sources, but also from foundations, patrons, and corporations (Bédécarrats et al., 2019; de Souza Leão & Eyal, 2019; Donovan, 2018). Indeed, as one interviewee explains, J-PAL sees itself largely "as a convener between donors who are interested in supporting [RCTs] and researchers who want to engage in the work" (Interview J-PAL 1). Establishing links with new funders is part of senior staff's everyday business. Describing J-PAL's fundraising efforts,

one annual report describes how its Executive Director, Rachel Glennerster, "presented at an Effective Altruism conference in California and had several follow-on meetings with potential donors from Silicon Valley", an audience from which J-PAL hoped to raise "up to US $10 million annually" (J-PAL, 2016, p. 22). In a follow-up to its influential 2006 report, the Center for Global Development (2022, p. 22) stresses that strengthening existing funding relations and establishing new ones is key for keeping up the momentum of evidence-based policymaking. Once again, RCTs seem to function as a hinge between academics and funders, promising dual rewards for both parties.

But again, the hinge turns out to be squeaky. As the Arnold Ventures researcher describes:

> *Arnold Ventures*: As funders, we didn't want to fund a bunch of beautiful studies that all came up with null findings. Which, it turns out, the [US] Department of Education, that's largely what happened there. They funded a ton of really great studies, but one after another they came back with disappointing findings. Because that can suck the life out of anything, you know, if you've got this great method [of RCTs] and you're finding out all these things that don't work. I mean, what's the path to improving people's lives then? So we decided that we were going to only fund trials where there was prior promising evidence.

This excerpt demonstrates at least two tensions in the hinge between the field of science, on the one hand, and the fields of business and politics on the other. First, while from a research-focused perspective finding out that programs do in fact *not* have the intended effects is just as valuable as finding out that they do, in the business perspective of Arnold Ventures a "null finding" is an obstacle to the evidence agenda. Considering that they personally support the program being tested and that their own money is at stake, demonstrating positive results is a much higher priority for funders than for researchers and evaluators. As Robert Granger, the former president of the William T. Grant Foundation notes, philanthropic foundations in the United States are becoming increasingly worried about "a cascade of mixed or null findings from Obama-era efforts" and "a restive practitioner community that has not seen strong benefits from rigorous evaluations" (Granger, 2018, pp. 151–152). Put more strongly, from a funder's perspective "rigorous evidence" could turn out to be self-defeating: if you seriously commit to RCTs, negative results threaten to disregard your pet policy—so do you really want to take chances? Indeed, generations of RCT advocates have repeatedly run into this very misalignment between scientific and applied fields (Campbell, 1969, pp. 409–410; Pritchett, 2002).

The second tension, as Ravallion (2020, p. 64) points out, is that Arnold Ventures' decision to "only fund trials where there was prior promising evidence" is in direct contradiction with the argument that a trial is ethical only if there is no *ex ante* evidence that a program has positive effects, a principle known as "equipoise" (MacKay, 2018). Among the researchers interviewed, agreement with this ethical proviso is virtually universal. Yet the incentives of funders point exactly in the opposite direction. For them, doing an RCT based on equipoise is the equivalent of throwing money out the window.

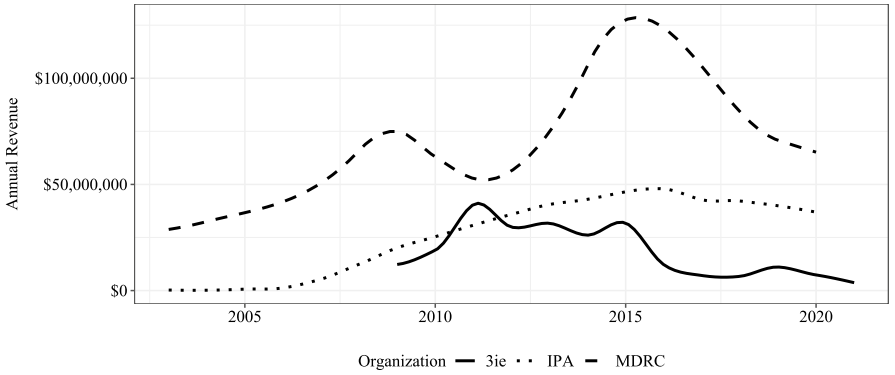

Fig. 3 Annual revenue of RCT supporting organizations over time. Data: Organizations' annual financial reports

There is some evidence that the tension between researchers' desire to accumulate evidence and funders' rationale to demonstrate positive results has become stronger over time. As one interviewee notes,

> *3ie*: In 2008, you had an environment where funders, particularly foundations like Gates and Hewlett, were willing to put money into the global public good of evidence. That was no longer true by 2015. So, the funding environment changed, people wanted back to... they really wanted things of interest to *them*, not global public goods. Even Gates threw a lot of money initially into 3ie in the first couple of years, but within a year and a half they were saying, "We wouldn't have done that now, we wouldn't give it now". And the money we got after that was for the particular grant programs they were interested in. But simple core funding to 3ie? That's gone.

It should be noted, however, that 3ie's experience seems to be relatively unique. As shown in Fig. 3, the revenue of RCT supporters (for whom numbers were available), peaked around 2015. Yet in the years that followed contributions did not decrease as much for other organizations as they did for 3ie.

Even so, RCT proponents have recently cautioned that "the financing of IEs [i.e., impact evaluations, which here means mostly RCTs] depends to a troubling extent on a small body of official agencies and foundations that regard IEs as extremely important products. Major shifts in policy by even a few such agencies could radically reduce the number of IEs being financed" (R. Manning et al., 2020, p. 38). Numerous interviewees worry about this possibility, complaining that RCT funders "wax and wane" in their commitment (Interview economist 2) or commenting that "it kind of goes in and out—sometimes philanthropies are more interested in evidence building, sometimes they become less interested because an advocacy agenda seems more important" (Interview MDRC).

Overall, this section demonstrates that the hinges RCTs have created between researchers and practitioners are weaker than usually thought. Researchers and policy-makers experience a constant tension between creating academic publications

and improving real-world policies and programs. Researchers and funders, for their part, are misaligned regarding the desire to learn new things and the rationale to invest in success stories. From this perspective, the progress of RCT proponents' self-proclaimed fight against ideology, ignorance, and inertia depends on their ability to manage the consonances, compromises, and contradictions of global interstitial fields.

Conclusion: The future of RCTs

This article has shown how support in favor of RCTs has crystallized into a global field of its own. Capitalizing on favorable institutional conditions in Northern and Southern countries, by the early 2000s many governments, NGOs, international organizations, research institutes, and philanthropic foundations found themselves in a situation in which conducting, funding, and collaborating on RCTs began to become a rewarding endeavor. To some extent, this "hinge" enabled researchers to better engage in research widely seen as especially rigorous, and it enabled practitioners to be seen as taking an evidence-based approach to policymaking.

At the same time, the article has shown that hinges among formerly separate fields have led to imperfectly aligned incentives. Because RCT supporters originate from the fields of science, politics, and business, they partly "inherit" these fields' stakes (or what field theorists call *illusio*). This, in turn, makes the global field suffer from goal conflicts among researchers, political practitioners, and funders. As an interstitial field, sitting in between more established fields, the global field of RCT support is therefore less stable than usually assumed. This analysis can be read as contribution to theoretical questions some social scientists are interested in: How do social fields emerge? How do they hang together? How do they influence each other? As I have suggested, the RCT story points to a more general hypothesis, in that the liberation from the constraints and expectations of more established fields that makes interstitial fields strong may be precisely what makes them weak. But the analysis also has wider, more practical implications. As I discuss now, assessing the effects of the interstitial field's internal dynamics and the critiques waged against RCTs leads to two predictions about the future of RCTs.

The first prediction is that the global field will probably differentiate according to the scientific, political, and economic fault lines that are observable already. This may imply professionalization and larger-scale trials for evaluations in applied policy contexts and a simultaneous tendency toward more technical "mechanism experiments" in more academic contexts (Ludwig et al., 2011). The former will probably be run by large firms specialized in RCTs, while the latter are run by academics. The official rationale may be a more productive distribution of labor, but in practice, research firms and academics need not have much to do with each other.

The basis of this prediction is the relative internal weakness of the global interstitial field of RCT support, discussed in this article, and the fact that the trends being predicted have already started to emerge. To begin with, differentiation into academic and applied branches would resolve some of the goal conflicts of the global interstitial field. Academics can publish slightly esoteric econometrics in specialized

journals without immediate pressure for "policy relevance". Specialized firms can run RCTs according to the preferences of their customers and make a profit, perhaps with the more business-savvy philanthropic foundations as investors. And political practitioners can consult with these firms, without having to debate researchers' concern for intellectual progress. This way, everyone is again aligned with the incentives of their respective fields. Constantly "lubricating the trust machine", as one researcher calls his task to maintain positive relations with practitioners (quoted in Fels, 2022, p. 24),[4] becomes obsolete—and so does the need to constantly lubricate the squeaky hinges between scientific and applied fields.

Initial evidence for the differentiation prediction comes from direct observations of RCT supporters as well as the historical development of adjacent fields. First, the only supporters of RCTs who seem largely exempt from the contradictory incentives of the interstitial field are, indeed, commercial firms. Openly acknowledging that they prioritize lucrative government contracts over academic publications, organizations like BIT, MDRC, or iNudgeYou can conduct academically uninteresting RCTs and replications without much concern for scientific recognition (Fels, 2022, pp. 22–26). As one BIT researcher puts it,

> *BIT*: We are a consultancy firm—we do what the customer wants. […] Sometimes, there are project teams who have the motivation to sit down to work on an academic publication. But that usually isn't the priority of our partners, our customers. So we must see how to finance such research papers for ourselves. […] Yet, because of this setup, as an organization we don't have this kind of conflict [between the interests of customers and academics].

In addition, differentiation into academic and applied RCTs is actually a well-known phenomenon in medicine and social science. Long having relied on academic specialists, by the 1990s tests of medical drugs came to be increasingly conducted by Contract Research Organizations. This became necessary because trials had to become larger, faster, and satisfy higher standards (Petryna, 2009). The development of market research and corporate data analysis is another example. Exempted from academic pressures to be "innovative", but equipped with large amounts of capital and specialized staff, such firms are the only actors who can pull off a large-scale population survey in a rigorous and timely fashion (Savage & Burrows, 2007; Vogel, 2019). Organizations that bridge academic and applied contexts—as J-PAL and IPA do now for policy trials—do not fit this differentiated environment. A corollary of the differentiation scenario is thus that such organizations would probably need to adapt their approach or find another niche.

The second prediction about the future of RCTs is that, as a method of policy evaluation, they will cease to be considered the "gold standard" and come to be seen

[4] The researcher in question, Dan Ariely, an expert on the psychology and economics of dishonesty, happens to be accused of "lubricating" not only his relations with practitioners but also his research data. This is not only very ironic, but it may also suggest that people frequently transitioning between fields have a harder time aligning values and priorities like "searching for the truth" and "generating maximum attention".

as one good approach among others. The hype will subside, extreme positions will become less common, and qualitative and quantitative approaches will again come closer together. In some sense, this prediction is connected to the first. But again, recent trends account for it independently.

Over the past few years, supporters of RCTs have adapted to several critiques raised against their favorite technique (see Bédécarrats et al., 2020; Deaton & Cartwright, 2018; Teele, 2014). The RCT proponents interviewed for this study almost universally acknowledge that experimental trials must be used selectively and cannot answer key macroeconomic questions regarding growth and inequality. Already in 2012, a study by the British Department for International Development (DFID, now FCDO), a key funder of RCTs, noted that "it is generally understood that methods and designs are fit for different purposes and when well-executed all have their strengths and weaknesses" (Stern et al., 2012, p. 9). Scholars have recently observed similar sentiments in Australia and France (Ball & Head, 2021; Devaux-Spatarakis, 2020). And as one J-PAL employee comments, RCTs are "a tool in the toolbox rather than the answer to all questions. It's very important. I think there needed to be a shift in framing and a shift in thinking, and that happened" (interview J-PAL 3).

Proponents also increasingly acknowledge concerns about equity, particularly regarding the enormous influence of researchers based in the Global North (CGD, 2022, pp. 21–22). As one recent study, published by a group of social experimenters, states,

> the vast majority of IE [i.e. impact evaluations, including RCTs] in LMICs [i.e. low- and middle-income countries] appear to have "northern" principal investigators. Undoubtedly, quality and rigour are essential to IEs, but it is important that IEs should not be perceived as a supply-driven product of a limited number of high-level academic departments in, for the most part, Anglo-Saxon universities, sometimes mediated through specialist consultancy firms (R. Manning et al., 2020, p. 37).

RCT proponents also worry that the price of running randomized experiments is very high, effectively excluding many NGOs from being "evidence-based" in the sense presently propagated. This leads to a serious ongoing debate on the meaning of "evidence" and the degree to which RCTs should designate its "gold standard" (interview Results for America). Perhaps the only critique supporters of RCTs generally reject is that experimental approaches are unethical in principle.

The ironic result of these adaptations to critique is that precisely the organizations who had committed to a strong focus on RCTs first now are the first to acknowledge their downsides. For instance, researchers at the Millennium Challenge Corporation noticed "just how hard it is to design and execute programmes with integrated, rigorous impact evaluations", hence "adapting its approach from one that prioritises impact evaluation to one that is selective in how and when to use impact evaluation" (Sturdy et al., 2014, p. 442). The Office of Management and Budget, which during the early 2000s had operated

as the "quarterback of evidence-based policy" in the United States, now argues that "a randomized controlled trial is not required for an evaluation to be rigorous, and using a method like a randomized controlled trial does not automatically ensure that an evaluation is conducted with the necessary rigor" (OMB, 2021, p. 11). The Behavioural Insights Team experienced "a necessity to move on, to innovate further—our focus now is much more on qualitative methods, on process evaluations" (interview BIT). And even according to one J-PAL researcher, "We're on the other end of the hump, where there was a peak in RCT interest and now there's actually a lot of revision of that interest" (interview J-PAL 2). All this supports the prediction that adaptation to ongoing criticism is slowly leading to a "new middle ground", reconnecting the "well-rehearsed and polarized positions" of experimental and non-experimental research (Gisselquist & Niño-Zarazúa, 2015, p. 2).

In both of the predicted scenarios, RCTs no longer function as hinges that hold together a "scientific business model" in which researchers and practitioners try to collaborate as best they can. Differentiation and adaptation to critique promise a more pragmatic (and hopefully more productive) use of RCTs. This pragmatic middle ground position may acknowledge that RCTs can answer some questions of social science, and they can answer some questions of public policy. But they have a hard time answering scientific *and* policy questions at the same time, meaning that different sorts of RCTs in different contexts are required. And they also have a hard time answering questions that are politically charged, ill-defined, or operating at the macro-level of social organization—in other words, lots of things—meaning that other techniques and approaches remain highly important. As science and society evolve, their methods of inquiry evolve with them. But tackling all problems with the same tool seems not only unwise—as a purely empirical matter, sustaining such an effort is also highly implausible.

Acknowledgments Research for this article was conducted during my time with the Research Training Group "World Politics". My colleagues' comments were usually bemused but often helpful. Can David Tobias and Frank Meyhöfer read the full manuscript and tried to talk me out of several crude ideas, metaphors, and analogies. Their arguments were strong, their dedication laudable, and their success partial.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Foundation (DFG), as a part of collective funding for the Research Training Group "World Politics: The emergence of political arenas and modes of observation in world society".

Data availability Data used for this article are available on request.

Declarations

Informed consent All people interviewed for this study were made aware of the potential risks and benefits of sharing information about their work. All gave their informed consent.

Competing interests The author has no competing interests to declare that are relevant to the content of this article.

References

Abbott, A. (2005). Linked ecologies: States and universities as environments for professions. *Sociological Theory, 23*(3), 245–274. https://doi.org/10.1111/j.0735-2751.2005.00253.x

Allen, T., & Parker, M. (2016). Deworming delusions? Mass drug administration in east African schools. *Journal of Biosocial Science, 48*(1), 116–147. https://doi.org/10.1017/S0021932016000171

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

AUE, & Nesta. (2011). *Using evidence to improve social policy and practice perspectives on how research and evidence can influence decision making*. https://media.nesta.org.uk/documents/using_evidence_to_improve_social_policy_and_practice.pdf. Accessed 6 June 2024

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology, 43*(1), 41–73. https://doi.org/10.1146/annurev-soc-073014-112445

Ball, S., & Head, B. W. (2021). Behavioural insights teams in practice: Nudge missions and methods on trial. *Policy & Politics, 49*(1), 105–120. https://doi.org/10.1332/030557320X15840777045205

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics,1*(1), 151–178. https://doi.org/10.1146/annurev.economics.050708.143235

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

Banerjee, A. V., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics,7*(1), 1–21. https://doi.org/10.1257/app.20140287

Bédécarrats, F., Guérin, I., & Roubaud, F. (2019). All that glitters is not gold. The political economy of randomized evaluations in development. *Development and Change, 50*(3), 735–762. https://doi.org/10.1111/dech.12378

Bédécarrats, F., Guérin, I., & Roubaud, F. (2020). Microfinance RCTs in Development: Miracle or Mirage? In F. Bédécarrats, I. Guérin, & F. Roubaud, *Randomized Control Trials in the Field of Development* (pp. 186–226). Oxford University Press. https://doi.org/10.1093/oso/9780198865360.003.0009

Berman, E. P. (2022). *Thinking like an economist: How efficiency replaced equality in U.S. Public Policy*. Princeton University Press.

Berndt, C. (2015). Behavioural economics, experimentalism and the marketization of development. *Economy and Society, 44*(4), 567–591. https://doi.org/10.1080/03085147.2015.1043794

BIT. (2012). *Behavioural insights team. Annual update 2011–2012*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/83719/Behavioural-Insights-Team-Annual-Update-2011-12_0.pdf. Accessed 6 June 2024

Bourdieu, P. (1985). The social space and the genesis of groups. *Theory and Society, 14*(6), 723–744.

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology*. Polity Press.

Breslau, D. (1998). *In search of the unequivocal. The political economy of measurement in U.S. Labor Market Policy*. Praeger.

Buchholz, L. (2016). What is a global field? Theorizing fields beyond the nation-state. In J. Go & M. Krause (Eds.), *Fielding Transnationalism* (pp. 31–60). Wiley Blackwell. http://doi.wiley.com/10.1002/2059-7932.12001

Cameron, D. (2010). *The next age of government*. https://www.youtube.com/watch?v=3ELnyoso6vI. Accessed 6 June 2024

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*(4), 409–429.

Carpenter, D. (2014). Reputation and power: Organizational image and pharmaceutical regulation at the FDA. *Princeton University Press*. https://doi.org/10.1515/9781400835119

CGD. (2006). *When will we ever learn? Improving lives through impact evaluation*. Center for Global Development.

CGD. (2022). *Breakthrough to policy use. Reinvigorating impact evaluation for global development*. Center for Global Development.

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). NBER Working Paper Series, Working Paper 25451. https://www.nber.org/system/files/working_papers/w25451/w25451.pdf. Accessed 6 June 2024

Daly, J. (2005). *Evidence-based medicine and the search for a science of clinical care*. University of California Press.

Davies, H. T. O., Nutley, S. M., & Smith, P. C. (2000). *What works? Evidence-based policy and practice in public services*. The Policy Press.

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: Reflections from England's education endowment foundation. *Educational Research, 60*(3), 292–310. https://doi.org/10.1080/00131881.2018.1500079

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. https://doi.org/10.1007/s11186-019-09352-6

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

Devaux-Spatarakis, A. (2020). Behind the scenes of the French evidence-based policy movement: Rise and fall of an RCT-driven model of evidence. In M. Palenberg & A. Paulson (Eds.), *The realpolitik of evaluation* (pp. 24–41). Routledge.

DIME. (2010). *Development impact evaluation initiative: A world bank-wide strategic approach to enhance developmental effectiveness*. World Bank. https://documents1.worldbank.org/curated/en/689141468161083406/pdf/633220WP0DIME000Box0361513B0PUBLIC0.pdf. Accessed 6 June 2024

Donaldson, S., Patton, M., Fetterman, D., and Scriven, M. (2010). The 2009 Claremont debates: The promise and pitfalls of utilization-focused and empowerment evaluation. CGU Faculty Publications and Research. https://scholarship.claremont.edu/cgu_fac_pub/563. Accessed 6 June 2024

Donovan, K. P. (2018). The rise of the Randomistas: On the experimental turn in international aid. *Economy and Society, 47*(1), 27–58.

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. In G. K. Pitman, O. N. Feinstein, & G. K. Ingram (Eds.), *Evaluating Development Effectiveness* (pp. 93–120). Transaction Publishers.

Dunleavy, P., & Hood, C. (1994). From old public administration to new public management. *Public Money & Management, 14*(3), 9–16. https://doi.org/10.1080/09540969409387823

Einfeld, C. (2019). Nudge and evidence based policy: Fertile ground. *Evidence & Policy, 15*(4), 509–524. https://doi.org/10.1332/174426418X15314036559759

Eyal, G. (2013). Spaces between fields. In P. S. Gorski (Ed.), *Bourdieu and historical analysis* (pp. 158–182). Duke University Press.

Fejerskov, A. (2022). *The global lab: Inequality, technology, and the experimental movement*. Oxford University Press.

Fels, K. M. (2022). Who nudges whom? Expert opinions on behavioural field experiments with public partners. *Behavioural Public Policy, 1–37*, 1. https://doi.org/10.1017/bpp.2022.14

Gërxhani, K., & Miller, L. (2022). Experimental sociology. In K. Gërxhani, N. de Graaf & W. Raub (Eds.), *Handbook of sociological science: contributions to rigorous sociology* (pp. 304–323). Edward Elgar Publishing. https://doi.org/10.4337/9781789909432

Gisselquist, R. M., & Niño-Zarazúa, M. (2015). What can experiments tell us about how to improve government performance? *Journal of Globalization and Development, 6*(1), 1–45. https://doi.org/10.1515/jgd-2014-0011

Granger, R. C. (2018). The roles foundations are playing in the evidence-based policy movement. *The ANNALS of the American Academy of Political and Social Science, 678*(1), 145–154. https://doi.org/10.1177/0002716218768801

Gueron, J. M. (2017). The politics and practice of social experiments: Seeds of a revolution. In A. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (pp. 27–69). Elsevier. https://doi.org/10.1016/bs.hefe.2016.11.001

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. Russell Sage Foundation.

Halpern, D. (2015). *Inside the nudge unit: How small changes can make a big difference*. Random House.

Halpern, D., & Mason, D. (2015). Radical incrementalism. *Evaluation, 21*(2), 143–149. https://doi.org/10.1177/1356389015578895

Harvey, C., Camasso, M. J., & Jagannathan, R. (2000). Evaluating welfare reform waivers under section 1115. *Journal of Economic Perspectives, 14*(4), 165–188. https://doi.org/10.1257/jep.14.4.165

Haskins, R., & Baron, J. (2011). *Building the connection between policy and evidence. The Obama evidence-based initiatives*.https://www.brookings.edu/wp-content/uploads/2016/06/0907_evidence_based_policy_haskins.pdf. Accessed 6 June 2024

Haskins, R., & Margolis, G. (2015). *Show me the evidence: Obama's fight for rigor and evidence in social policy*. Brookings Institution Press.

Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research, 60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Hook, S. W. (2008). Ideas and change in U.S. foreign aid: Inventing the millennium challenge corporation. *Foreign Policy Analysis, 4*(2), 147–167. https://doi.org/10.1111/j.1743-8594.2007.00062.x

Humphreys, M., & Weinstein, J. M. (2009). Field experiments and the political economy of development. *Annual Review of Political Science, 12*(1), 367–378. https://doi.org/10.1146/annurev.polisci.12.060107.155922

Jamison, J. C. (2019). The entry of randomized assignment into the social sciences. *Journal of Causal Inference, 7*(1), 1–16. https://doi.org/10.1515/jci-2017-0025

Jatteau, A. (2018). The success of randomized controlled trials: A sociographical study of the rise of J-PAL to scientific excellence and influence. *Historical Social Research/Historische Sozialforschung*, *43*(3), 94–119. https://doi.org/10.12759/HSR.43.2018.3.94-119

John, P. (2018). *How Far to Nudge?* Assessing behavioural public policy. Edward Elgar.

Jones, R., & Whitehead, M. (2018). 'Politics done like science': Critical perspectives on psychological governance and the experimental state. *Environment and Planning d: Society and Space, 36*(2), 313–330. https://doi.org/10.1177/0263775817748330

J-PAL. (2016). *2016 Annual Report*. Massachusetts Institute of Technology.

Karlan, D. (2011). *More than good intentions: How a new economics is helping to solve global poverty*. Dutton.

Kelly, A. H., & McGoey, L. (2018). Facts, power and global evidence: A new empire of truth. *Economy and Society, 47*(1), 1–26. https://doi.org/10.1080/03085147.2018.1457261

Krause, M. (2018). How fields vary. *The British Journal of Sociology, 69*(1), 3–22. https://doi.org/10.1111/1468-4446.12258

Krause, M. (2021). *Model cases: On canonical research objects and sites*. The University of Chicago Press.

Labrousse, A. (2020). The rhetorical superiority of poor economics. In F. Bédécarrats, I. Guerin, & F. Roubaud (Eds.), *Randomized control trials in the field of development* (pp. 227–255). Oxford University Press. https://doi.org/10.1093/oso/9780198865360.003.0010

Lee, C., & Ma, L. (2020). The role of policy labs in policy experiment and knowledge transfer: A comparison across the UK, Denmark, and Singapore. *Journal of Comparative Policy Analysis: Research and Practice, 22*(4), 281–297. https://doi.org/10.1080/13876988.2019.1668657

Leigh, A. (2018). *Randomistas: How radical researchers are changing our world*. Yale University Press.

Liu, S. (2021). Between social spaces. *European Journal of Social Theory, 24*(1), 123–139. https://doi.org/10.1177/1368431020905258

European Union (2016). *Behavioural insights applied to policy. European report 2016*. European Commission.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives, 25*(3), 17–38. https://doi.org/10.1257/jep.25.3.17

MacKay, D. (2018). The ethics of public policy RCTs: The principle of policy equipoise. *Bioethics, 32*(1), 59–67. https://doi.org/10.1111/bioe.12403

Manning, N. (2001). The legacy of the new public management in developing countries. *International Review of Administrative Sciences, 67*(2), 297–312. https://doi.org/10.1177/0020852301672009

Manning, R., Goldman, I., & Hernández Licona, G. (2020). The impact of impact evaluation: Are impact evaluation and impact evaluation synthesis contributing to evidence generation and use in low- and

middle-income countries? *WIDER Working Paper 2020/20*, *2020*. https://doi.org/10.35188/UNU-WIDER/2020/777-4

Marks, H. M. (2000). *The progress of experiment: Science and therapeutic reform in the United States, 1900–1990*. Cambridge University Press.

Medvetz, T. (2012). *Think tanks in America*. University of Chicago Press.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Moynihan, D. P. (2013). Advancing the empirical study of performance management: What we learned from the program assessment rating tool. *The American Review of Public Administration, 43*(5), 499–517. https://doi.org/10.1177/0275074013487023

Neuwinger, M. (2023). Are social experiments being hyped (too much)? *Journal for Technology Assessment in Theory and Practice, 32*(3), 22–27. https://doi.org/10.14512/tatup.32.3.22

Nussle, J., & Orszag, P. (2015). *Moneyball for government* (2nd ed.). Disruption Books.

OECD. (2017). *Behavioural insights and public policy: Lessons from around the world*. OECD Publishing.

Oliver, K. (2022). How policy appetites shape, and are shaped by evidence production and use. In P. Fafard, A. Cassola, & E. de Leeuw (Eds.), *Integrating Science and Politics for Public Health* (pp. 77–101). Springer. https://doi.org/10.1007/978-3-030-98985-9_5

OMB. (2004). *What constitutes strong evidence of program effectiveness*. https://obamawhitehouse.archives.gov/sites/default/files/omb/part/2004_program_eval.pdf. Accessed 6 June 2024

OMB. (2021). *Memorandum for Heads of Executive Departments and Agencies*. https://www.whitehouse.gov/wp-content/uploads/2021/06/M-21-27.pdf. Accessed 6 June 2024

Orr, L. L. (2018). The role of evaluation in building evidence-based policy. *The ANNALS of the American Academy of Political and Social Science, 678*(1), 51–59. https://doi.org/10.1177/0002716218764299

Page, S. (2005). What's new about the new public management? Administrative change in the human services. *Public Administration Review, 65*(6), 713–727. https://doi.org/10.1111/j.1540-6210.2005.00500.x

Pamies-Sumner, S. (2015). *Development impact evaluations: State of play and new challenges*. https://www.afd.fr/en/ressources/development-impact-evaluations-state-play-and-new-challenges. Accessed 6 June 2024

Parker, I. (2010). The Poverty Lab. *The New Yorker*. https://www.newyorker.com/magazine/2010/05/17/the-poverty-lab. Accessed 6 June 2024

Pearce, W., & Raman, S. (2014). The new Randomised Controlled Trials (RCT) movement in public policy: Challenges of epistemic governance. *Policy Sciences, 47*(4), 387–402. https://doi.org/10.1007/s11077-014-9208-3

Petryna, A. (2009). *When Experiments Travel*. Princeton University Press.

Picciotto, R. (2012). Experimentalism and development evaluation: Will the bubble burst? *Evaluation, 18*(2), 213–292. https://doi.org/10.1177/1356389012440915

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., & Andersen, S. C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research, 60*(3), 311–335. https://doi.org/10.1080/00131881.2018.1493351

Pritchett, L. (2002). It pays to be ignorant: A simple political economy of rigorous program evaluation. *Journal of Policy Reform, 5*(4), 251–269. https://doi.org/10.1080/1384128032000096832

Ravallion, M. (2020). Should the Randomistas (Continue to) Rule? In F. Bédécarrats, I. Guerin, & F. Roubaud (Eds.), *Randomized Control Trials in the Field of Development* (pp. 47–78). Oxford University Press. https://doi.org/10.1093/oso/9780198865360.003.0003

Rodrik, D. (2006). Goodbye Washington consensus, hello Washington confusion? A review of the world bank's economic growth in the 1990s: Learning from a decade of reform. *Journal of Economic Literature, XLIV*, 973–987.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology,41*(5), 885–899. https://doi.org/10.1177/0038038507080443

Schedler, K., & Proeller, I. (2002). The new public management: A perspective from mainland Europe. In K. McLaughlin, S. P. Osborne, & E. Ferlie (Eds.), *New public management* (pp. 163–180). Routledge.

Sent, E.-M. (2004). Behavioral economics: How psychology made its (limited) way back into economics. *History of Political Economy, 36*(4), 735–760. https://doi.org/10.1215/00182702-36-4-735

Stack, K. (2018). The office of management and budget: The quarterback of evidence-based policy in the federal government. *The ANNALS of the American Academy of Political and Social Science, 678*(1), 112–123. https://doi.org/10.1177/0002716218768440

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations*. Institute for Development Studies. http://repository.fteval.at/id/eprint/126. Accessed 6 June 2024

Sturdy, J., Aquino, S., & Molyneaux, J. (2014). Learning from evaluation at the millennium challenge corporation. *Journal of Development Effectiveness, 6*(4), 436–450. https://doi.org/10.1080/19439342.2014.975424

Teele, D. L. (2014). Reflections on the Ethics of Field Experiments. In D. L. Teele (Ed.), *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* (pp. 115–140). Yale University Press.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review, 93*(3), 175–179.

Thornton, P. H., & Ocasio, W. (2008). Institutional Logics. In R. Greenwood, C. Oliver, R. Suddaby & K. Sahlin (Eds.), *The SAGE Handbook of Organizational Institutionalism* (pp. 99–128). Sage. https://doi.org/10.4135/9781849200387.n4

Tollefson, J. (2015). Revolt of the Randomistas. *Nature, 524*, 150–153.

United Nations. (2016). *Behavioural insights at the United Nations.* Achieving Agenda 2030. UN.

Vedung, E. (2010). Four waves of evaluation diffusion. *Evaluation, 16*(3), 263–277. https://doi.org/10.1177/1356389010372452

Vogel, R. (2019). *Survey-Welten: Eine empirische Perspektive auf Qualitätskonventionen und Praxisformen der Umfrageforschung*. Springer. https://doi.org/10.1007/978-3-658-25437-7

Wacquant, L. (2019). Bourdieu's Dyad: On the primacy of social space and symbolic power. In J. Blasius, F. Lebaron, B. Le Roux, & A. Schmitz (Eds.), *Empirical Investigations of Social Space* (pp. 15–21). Springer.

Wacquant, L., & Akçaoğlu, A. (2017). Practice and symbolic power in Bourdieu: The view from Berkeley. *Journal of Classical Sociology, 17*(1), 55–69. https://doi.org/10.1177/1468795X16682145

White, H. (2014). Current challenges in impact evaluation. *The European Journal of Development Research, 26*(1), 18–30. https://doi.org/10.1057/ejdr.2013.45

White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Humanities & Social Sciences Communications, 5*(1), 47. https://doi.org/10.1057/s41599-019-0253-6

Whitehead, M., Jones, R., Howell, R., Lilley, R., & Pycket, J. (2014). *Nudging all over the world. Assessing the global impact of the behavioural sciences on public policy*. Aberystwyth University. https://changingbehaviours.files.wordpress.com/2014/09/nudgedesignfinal.pdf. Accessed 6 June 2024

Whitehead, M., Jones, R., Lilley, R., Pycket, J., & Howell, R. (2018). *Neuroliberalism: Behavioural government in the twenty first century*. Routledge.

Whitehurst, G. J. (Russ). (2018). The institute of education sciences: A model for federal research offices. *The ANNALS of the American Academy of Political and Social Science, 678*(1), 124–133. https://doi.org/10.1177/0002716218768243

Williams, J. W. (2023). "Let's not have the perfect be the enemy of the good": Social impact bonds, randomized controlled trials, and the valuation of social programs. *Science, Technology, & Human Values, 48*(1), 91–114.

World Bank. (2015). World development report 2015: Mind, society, and behavior. *The World Bank*. https://doi.org/10.1596/978-1-4648-0342-0

World Food Programme. (2019). *WFP Impact Evaluation Strategy (2019—2026)*. WFP Office of Evaluation. https://docs.wfp.org/api/documents/WFP-0000109085/download/. Accessed 6 June 2024

Malte Neuwinger is a doctoral researcher at Bielefeld University.