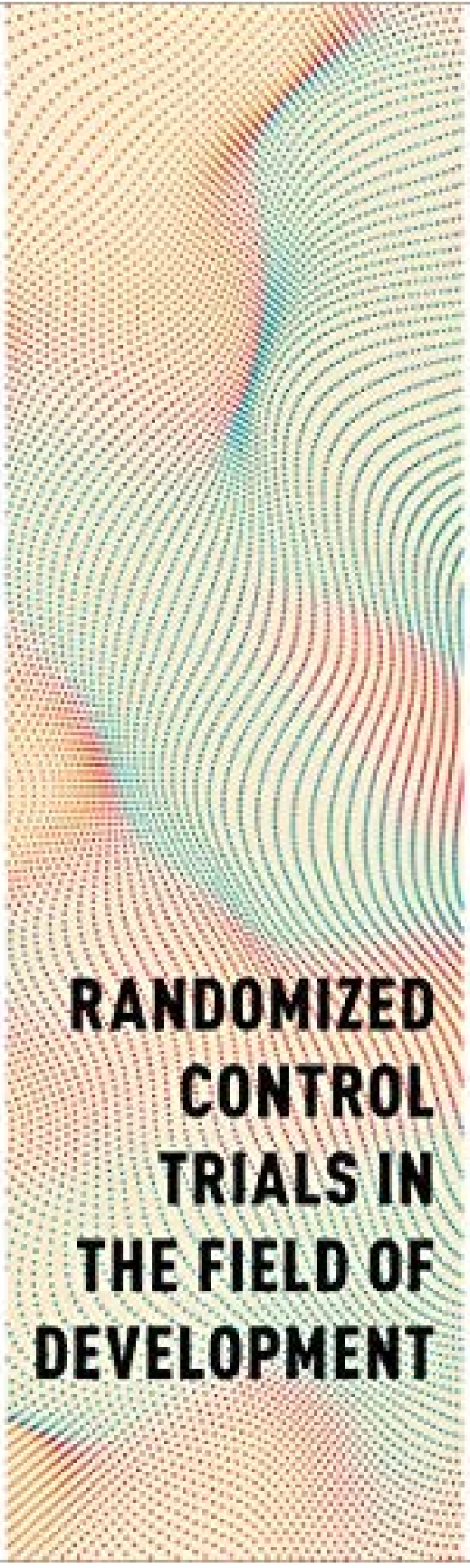

OXFORD

RANDOMIZED CONTROL TRIALS IN THE FIELD OF DEVELOPMENT

a critical perspective

edited by

Florent Bédécarrats

Isabelle Guérin

François Roubaud

Randomized Control Trials in the Field of Development

A Critical Perspective

Edited by

FLORENT BÉDÉCARRATS
ISABELLE GUÉRIN
FRANÇOIS ROUBAUD

OXFORD
UNIVERSITY PRESS
2020

Contents

List of Illustrations

List of Tables

Acknowledgments

The idea for this collective volume was born in the summer of 2018, with the desire to foster a scientific controversy about a method whose influence was growing and, in our opinion, remained insufficiently challenged. The nobelization of Abijit Banerjee, Esther Duflo, and Michael Kremer took place just as we were finalizing the manuscript: it makes the controversy all the more urgent and necessary.

To implement this project, we did not receive any particular "project" or "funding", which proves that it is still possible to do research in a free and disinterested way.

Nevertheless, we did obtain one occasional financial support from our home institutions—AFD and IRD—to organize a seminar bringing together most of the authors, which took place in Paris in March 2019. This seminar allowed us to discuss and debate, including our disagreements, since the purpose of this book is not to converge on everything, nor to propose a "ready-to-think" on how to approach RCTs, but to set out the terms of the debate and the foundations for the controversy.

We therefore thank all those (institutions or individuals) who have agreed to bring to life, each in his or her own way, this controversy that has been avoided for too long: AFD and IRD for their support in this event, and their respective CEOs who courageously signed an afterword to the book; Gaël Giraud, Chief Economist of AFD at the time we took the initiative for this book, and who has been very supportive; all the authors of this book, in their diversity and the plurality of their positions, for believing in our proposal, for having played the game of controversy by giving the best of themselves and for their patience; Lant Pritchett, who put us in contact with Gulzar Natarajan and Ila Patnaik: their testimonies at the end of the book shed light on the field and the decision-maker's point of view, a necessary complement to the researcher's viewpoint; Britta Augsburg, who agreed to come and debate with us on the substance, when other RCT proponents declined our invitations; Diane Bertrand, our historical translator, who once again worked miracles (but perhaps it is a mirage?); and lastly, our publisher, Oxford University Press, for its commitment, as a guarantor of academic freedom, at our side.

Let's bet that this book will contribute to fuel the flow of ideas, science in action; let our future readers confirm it.

Notes on Contributors

Michel Abramowicz is a Cardiologist affiliated with the Erasme Hospital, the academic hospital of the Université Libre de Bruxelles (ULB), Belgium, where he oversaw the multidisciplinary seminar for four years. He published articles and letters in scientific journals such as the *American Journal of Cardiology*, the *American Journal of Respiratory and Critical Care Medicine*, the *American Journal of Epidemiology,* the *European Heart Journal*, and the *New England Journal of Medicine*. While Head of the Coronary Care Unit, he was actively involved as a field investigator in the Second International Study of Infarct Survival (ISIS-2), one of the first cardiological mega-RCTs.

Radu Ban is Senior Program Officer, Water, Sanitation & Hygiene Program at the Bill & Melinda Gates Foundation. He is leading the Measurement and Evidence work on the Water, Sanitation & Hygiene Program. In his position, he manages a research portfolio to better understand what approaches for sanitation service delivery work and how. Radu Ban is a development economist by training, and obtained his PhD from the London School of Economics. Prior to joining the foundation, he worked as an economist at the World Bank's Development Impact Evaluation (DIME) initiative, focusing on governance and community-driven development.

Florent Bédécarrats holds a PhD from the University of Paris-Sorbonne. Since late 2019, he has been heading the data management unit at Nantes Metropole. Most of his work on this book was done while being in charge of coordinating scientific impact evaluations at the French Development Agency. He held this position from 2013 to 2019. From 2007 to 2013 he was in charge of research and development activities at CERISE, a platform of microfinance support organizations. Previously, he worked for three years in Latin America, in a solidarity based company for tourism and culture in Brazil, for a network of microfinance cooperatives in Mexico, and for an international NGO in Guatemala.

Matthew H. Bonds is assistant professor at Harvard Medical School. Dr Bonds has a PhD in economics and a PhD in ecology from the University of Georgia. He joined the Harvard Medical School faculty after a postdoctoral fellowship in sustainable development under the mentorship of Jeffrey Sachs at the Earth Institute at Columbia University. While developing formal theoretical frameworks on poverty traps, Dr. Bonds has worked with Partners In Health in Rwanda and is the co-founder and co-CEO of the healthcare NGO PIVOT in Madagascar. His research focus is on building new evaluation methods in global health.

Oliver Cumming is an Assistant Professor at the London School of Hygiene and Tropical Medicine and the Deputy Director of the Environmental Health Group. He is currently the Principal Investigator for studies in several countries, including Mozambique, Kenya, Senegal, and the Democratic Republic of Congo, and serves as the Research Director of the SHARE Consortium. His research focuses on the epidemiology of water, sanitation, and hygiene (WASH)-related diseases, and he currently leads multiple trials to evaluate the

effect of different WASH interventions on various child health outcomes, including: enteric infections, child growth and development, and oral vaccine performance.

Angus Deaton is Professor Emeritus at Princeton University and Presidential Professor at USC. He is the author of *The Great Escape* and, with Anne Case, *Deaths of Despair and the Future of Capitalism*. He works on health, happiness, development, poverty, inequality, and evidence for policy. A member of the National Academy of Sciences, a Fellow of the British Academy, and an Honorary Fellow of the Royal Society of Edinburgh, in 2015, he received the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. He was born in Edinburgh, Scotland. He was made a Knight Bachelor in 2016.

Paul E. Farmer holds an MD and PhD from Harvard University, where he is the Kolokotrones University Professor and Chair of the Department of Global Health and Social Medicine at Harvard Medical School. He is Co-Founder and Chief Strategist of Partners In Health, an international non-profit organization that since 1987 has provided direct health care services and undertaken research and advocacy activities on behalf of those who are sick and living in poverty. He is Professor of Medicine and Chief of the Division of Global Health Equity at Brigham and Women's Hospital. Additionally, Dr. Farmer serves as the United Nations Special Adviser to the Secretary-General on Community Based Medicine and Lessons from Haiti.

Andres Garchitorena is a researcher at the French Research Institute for Sustainable Development (IRD). He holds a V.M.D and PhD in public health. He joined the IRD faculty after a postdoctoral fellowship in global health at Harvard Medical School. He is also associate scientific director at PIVOT, a health system strengthening organization with a strong research focus working in partnership with the Madagascar Ministry of Health, where he develops novel impact evaluations of the HSS intervention using population and health system data.

Isabelle Guérin, PhD, is a socioeconomist, Senior Research Fellow at the French Institute of Research for Sustainable Development (IRD), Associate at the French Institute of Pondicherry, and presently member of the School of Social Sciences at the Institute for Advanced Study, Princeton (2019–2020). She specializes in the political and moral economics of money, debt, and finance. Her current work focuses on the financialization of domestic economies, looking at how financialization produces new forms of inequalities and domination, but also alternative and solidarity-based initiatives. Her work draws most often from her own field-based original data, combines ethnography and statistical analyses, and is interdisciplinary and comparative in nature. Her work also includes a permanent thinking about the conditions of data production and the combination of methods.

James J. Heckman is the Henry Schultz Distinguished Service Professor of Economics and Public Policy at the University of Chicago. He works to understand the origins of inequality, and skill formation, and develops and applies strategies for addressing these issues. Heckman has published over 300 articles and nine books. Heckman received the Nobel Prize in Economics, Dan David Prize, and Chinese Government Friendship Award, among other recognitions. He is Director of the Center for the Economics of Human Development at the University of Chicago. The Center investigates the sources of poverty and social immobility and policies to improve human flourishing.

Bethany Hedt-Gauthier is associate professor at Harvard Medical School. She is a biostatistician notable for her innovative application, development, and evaluation of research methodologies toward improvement in the health of populations in resource-limited environments. Her current research focuses primarily on health systems strengthening in Africa, with a focus on global surgery. In addition, she leads work on equity in global health research collaborations with a focus on comprehensive training programs to build national capacity to support and lead research.

Agnès Labrousse is Associate Professor in Economics at the University of Picardie, France, and Associate Editor of the *Régulation Review*. She has been working on epistemology, the pharmaceutical industry, and development issues from an institutionalist perspective. Her work on RCTs bridges these areas of research and her article "Not by Technique Alone. Comparing Development Analysis with Elinor Ostrom and Esther Duflo" in the *Journal of Institutional Economics* received both the 2017 Ostrom Prize and the 2016 EAEPE Kapp-Prize.

Jean Paul Moatti is Emeritus Professor of Health Economics at Aix Marseille University (AMU) in South Eastern France and has extensively worked on Aids TB and malaria as well as access to essential medicines in developing countries. Between March 2015 and February 2020, he has been the CEO of the French National Research Institute for Sustainable Development (IRD), in charge of scientific partnership with developing countries. He has been a member of the Independant Group of Scientists that produced the first United Nations quadriennal evaluation report about the implementation of the Sustainable Development Goals (GSDR2019).

Jonathan Morduch is Professor of Public Policy and Economics at the Wagner Graduate School of Public Service at New York University. His research focuses on poverty, inequality, and finance. He is the author with Rachel Schneider of *The Financial Diaries: How American Families Cope in a World of Uncertainty* (Princeton 2017) and a co-author of *Portfolios of the Poor: How the World's Poor Live on $2 a Day* (Princeton 2009). Morduch has also co-written *The Economics of Microfinance* (MIT Press 2010); and *Economics* (McGraw-Hill 2017, 2nd ed.). He is a founder and Executive Director of the NYU Financial Access Initiative.

Megan B. Murray is professor of global health and social medicine in the Department of Global Health and Social Medicine and associate professor of medicine in the Department of Medicine, Harvard Medical School, and professor of epidemiology at Harvard School of Public Health. Dr. Murray directs the Department of Global Health and Social Medicine Research Core; she is also director of research for the Division of Global Health Equity at Brigham and Women's Hospital. Besides outstanding research experience in infectious disease epidemiology, her research focus includes health system strengthening impact evaluations in Rwanda and Madagascar.

Gulzar Natarajan is an officer of the Indian Administrative Service. Over a twenty years career, he has served in the office of the Prime Minister of India, managed the Infrastructure Corporation of the Andhra Pradesh state, been District Collector of Hyderabad, Chairman and Managing Director of a power distribution company based at Visakhapatnam, Municipal Commissioner of Vijayawada, and in development field postings across Andhra Pradesh. He has also led the design and implementation of large-scale projects in

infrastructure, urban, health, education, skills and livelihoods, poverty reduction etc. across various levels of the government. He holds a bachelor's Engineering degree from the Indian Institute of Technology, Chennai, and a master's degree in International Development (MPA-ID) from Harvard Kennedy School.

Timothy N. Ogden is Managing Director of the Financial Access Initiative at NYU-Wagner, and a senior fellow of the Aspen Institute's Economic Opportunities Program and Financial Security Program. He also serves as Executive Partner of Sona Partners, Chair of the Board of GiveWell, and President of the Bardet Biedl Syndrome Foundation. Ogden is the editor of the faiV, a widely read newsletter on financial inclusion, digital finance, evidence-based policy and economic development. His book, *Experimental Conversations: Perspectives on the Use of Randomized Trials in Development Economics*, collects interviews with 20 leading thinkers on the topic.

Ila Patnaik is a professor at the National Institute of Public Finance and Policy, New Delhi. Prior to this, she was the Principal Economic Advisor to the Government of India. Her research interests include international macroeconomics, finance, and emerging economy business cycles and financial sector regulation. She has publications in scholarly journals such as the *Journal of International Money and Finance*, *The World Bank Economic Review* and *International Finance*. Dr. Patnaik also served on various working groups and task forces of the Ministry of Finance.

Robert Picciotto, Adjunct Professor, University of Auckland, and Senior Independent Evaluation Adviser to the Ministry of Foreign Affairs and Trade in New Zealand, is a graduate of Princeton University and a member of the Academy of Social Sciences. He retired from the World Bank in 2002 after holding several operational and corporate management positions, including Vice-President, Corporate Planning and Budgeting and Director-General of the Independent Evaluation Group for two consecutive five-year terms.

Lant Pritchett is research director of the RISE project at Oxford's Blavatnik School of Government and an associate of the Building State Capability project at Harvard Kennedy School. After a PhD in Economics in 1983 from MIT he worked with the World Bank from 1988 to 2007, living in Indonesia 1998–2000 and India 2004–2007. He taught at Harvard Kennedy School between 2000 and 2018. He has over a hundred publications (with over fifty different co-authors) on a wide range of development topics (education, economic growth, state capability, labor mobility, poverty, and learning from RCTs).

Martin Ravallion currently holds the inaugural Edmond D. Villani Chair of Economics at Georgetown University. Prior to joining Georgetown in 2013 he was Director of the World Bank's research department, the Development Research Group. He joined the Bank in 1988 and worked in almost all sectors and all regions over the following 24 years. Prior to joining the Bank, Martin was on the faculty of the Australian National University. Martin's main research interests over the last 30 years have concerned poverty and policies for fighting it. He has advised numerous governments and international agencies on this topic.

Rémy Rioux is a Senior Advisor at the Court of Auditors. He is an expert in international financial institutions and has held high-level positions in a career devoted to development and Africa. After serving as Director of the Office of the Minister of the Economy, Finance and Foreign Trade, Pierre Moscovici, he was appointed Deputy General Secretary by

Laurent Fabius, Minister for Foreign Affairs and International Development, and coordinated the "finance" agenda for the French presidency of COP21. Since 2016, he has headed the French Development Agency (AFD). Rémy Rioux is also Chairman of the International Development Finance Club (IDFC), the largest provider of development and climate finance globally.

François Roubaud, PhD, is an economist and statistician, a senior research fellow at the French Institute of Research for Sustainable Development (IRD), a member of the DIAL research unit in Paris and former head (2000–2004). He holds a PhD in Economics from the Paris-Ouest Nanterre University and is a graduate of the Paris Graduate School of Economics, Statistics and Finance (ENSAE). In statistics, he initiated the mixed surveys approach (household-enterprise) to measure the informal economy, in particular the *1-2-3 survey*, and developed the governance modules grafted on official household surveys now used to monitor SDG16. Both are recognized as international standards and implemented in dozens of LDCs (in Africa, Latin America and Asia). In development economics, his main fields of expertise are labour market and informal economy, corruption, governance and institutions, and impact evaluation and political economic of development policies.

Dean Spears is the Executive Director of the Research Institute for Compassionate Economics (RICE). Dean is an economic demographer and development economist. His research areas include: the health, growth, and survival of children, especially in India; the environment, pollution, and climate change; and population dimensions of social well-being. In addition to being a founding Executive Director of RICE, Dean is Assistant Professor of Economics at the University of Texas at Austin, is a visiting economist at the Economic and Planning Unit of the Indian Statistical Institute in Delhi, and is an affiliate of the Climate Futures Initiative at Princeton University. With Diane Coffey, he is the author of the award-winning book *Where India Goes: Abandoned Toilets, Stunted Development, and the Costs of Caste.*

Ariane Szafarz is a Professor of finance at ULB, SBS-EM, Belgium, and a Co-Director of the Centre for European Research in Microfinance (CERMi). She holds a PhD in Mathematics and an MA in Philosophy of Science. She currently works on microfinance, social banking, mission drift, and gender discrimination. She has published several books and articles in, e.g., *Academy of Management Review, European Economic Review, Journal of Banking and Finance, Journal of Business Ethics, Journal of Development Studies, Journal of International Money and Finance, Review of Finance*, *World Development.* Two of her co-authored articles received the Warren Samuels Prize (2016 and 2019) awarded at the ASSA Meetings by the Association for Social Economics.

Eva Vivalt is a senior lecturer in economics at the Australian National University. Dr. Vivalt's main research interests are in investigating stumbling blocks to generating evidence-based policy decisions, including both methodological issues as well as how evidence is interpreted and used. Dr. Vivalt is also a Principal Investigator on Y Combinator Research's basic income RCT and has other interests in development, behavioral economics, and effective altruism. Dr. Vivalt has recently been working on forecasts of social science results and, with Stefano DellaVigna, built a platform researchers can use to elicit predictions for their own studies, the Social Science Prediction Platform.

Editors' Introduction: Controversies around RCT in Development

Epistemology, Ethics, and Politics

Florent Bédécarrats, Isabelle Guérin, and François Roubaud

In October 2019, Abhijit Banerjee, Esther Duflo, and Michael Kremer jointly won the 51st Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. The three researchers were awarded "for their experimental approach to alleviating global poverty" and for having "turned development economics—the field that studies what causes global poverty and how best to combat it—into a blossoming, largely experimental field" (The Royal Swedish Academy of Sciences 2019: 2). The use of field experimentation, unlike laboratory experimentation, serves to conduct full-scale tests on interventions, behaviour, and decision-making in the "real world," and then to "make causal claims of impact" (ibid., 3). Consequently, stated the jury, "We now have a large number of concrete results on specific mechanisms behind poverty and specific interventions to alleviate it" (ibid.). The cases of health, schooling, gender and politics, and credit are given as powerful illustrations of the laureates' achievements in their work. This award recognizes the success of a long-standing method inspired by the medical field—randomized control trials (hereinafter referred to as RCTs)—and now applied to poverty and development issues. The award did not really come as a surprise. RCTs were first launched in development in the early 2000s and have since become increasingly successful among academics, donors, and development practitioners to the extent that RCTs are now considered the gold standard for the evaluation of anti-poverty policies and understanding the origins of poverty.

While there are reasons to welcome the prize (one of the three laureates is a young woman,[1] and the award brings to the fore the issue of poverty and the collection of primary data, which has long been passed over by development economics), there is also cause to raise questions about the validity and repercussions of the growing use of this method, which the prize may boost further.

[1] The prize first awarded in 1969 has been won by a total of 84 laureates. Esther Duflo is only the second-ever female laureate. Over and above the prize itself, economics as a social science is the most marked by discrimination against women (Lundberg and Stearns 2019).

Florent Bédécarrats, Isabelle Guérin, and François Roubaud, *Editors' Introduction: Controversies around RCT in Development: Epistemology, Ethics, and Politics* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
 DOI: 10.1093/oso/9780198865360.003.0001

What scope do RCTs actually have? Have they really "dramatically improved our ability to fight poverty in practice," as suggested by the Sveriges Riksbank Prize jury? Which sorts of questions are RCTs able to address and which do they fail to answer? Is causal explanation the only way to understand poverty and do RCTs systematically manage to provide causal explanations? Last, but not least, is the supremacy of experimentation in development economics, as recognized and commended by the Nobel jury, scientifically legitimate and politically desirable?

This edited volume proposes to answer these questions. The initiative for this editorial project came from the EUDN[2] conference on *Malaise dans l'Evaluation* (Evaluation and its Discontents) held by AFD[3] in Paris in 2012 (AFD 2012). At this event, we witnessed a real dialogue of the deaf. Whereas some critical voices set out the reasons for their doubts, those who we will call the *randomistas*,[4] in keeping with others and for the sake of expediency, confidently presented their convictions and their findings, sidestepping any substantive discussion of the matter.

We therefore decided to analyze the success of RCTs, taking three angles (Bédécarrats, Guérin, and Roubaud 2013, 2019; Bédécarrats et al. 2019a, 2019b): developing theoretical critiques based on the classic internal and external validity questions (RCTs in theory: *doing the maths*); focusing the critique empirically: how RCTs are conducted on the ground (RCTs in practice: *doing the cooking*); and analyzing the political economy of RCTs in terms of both supply and demand (RCTs as a business: *doing the accounts, both financial and symbolic*). Whereas the first point had been largely explored and our contribution marginal, the other two matters were relatively uncharted territory.[5] Our own analyses come from an in-depth observation of two RCTs (microcredit in Morocco (Morvant-Roux et al. 2014) and micro-insurance in Cambodia (Quentin and Guérin 2013)) were largely borne out by an analysis of three of the most emblematic RCTs.[6] These RCTs ultimately

[2] European Development Network. [3] French Agency for Development.

[4] We mean by this term those researchers defending the superiority of the method over all others. On the non-pejorative term of "*randomista*," see the chapters by Ravallion (Chapter 1) and Ogden (Chapter 4). See also Gibson (2019).

[5] Two main conclusions emerged from our analyses. First, although RCTs represent a suitable way to estimate the causal impact of a certain number of bounded projects, this is only true in ideal conditions defined in theory and rarely observed on the ground. And in these ideal conditions, RCTs may be able to be used to statistically quantify the impacts (significance and magnitude), but they cannot identify the mechanisms through which these impacts channel (paradoxical for a method that makes the analysis of causality its fundamental principle). Second, three of the major claims made by randomistas are groundless: i.e. that RCTs are superior to any other method; that the proliferation of RCTs can solve the external validity issue, acknowledged by all as an intrinsic weakness (which we have termed a "hegemonic plan"); and that RCTs can provide all the answers when it comes to "what works and doesn't work in development."

[6] The famous RCT associated with the conditional cash transfer programme in Mexico (Progresa, renamed Oportunidades and then Prospera), which many see as the catalyst for the rush on RCTs, and CCTs accordingly, but whose implementation and hence internal validity are disputed (Faulkner 2014); the equally high-profile RCT on intestinal worms in Kenya by Miguel & Kremer (2004), whose findings have been challenged by a group of epidemiologists (Aiken et al. 2015; Davey et al. 2015; Humphreys 2015), which is paradoxical given that the *randomistas* have made RCTs in medicine the movement's flagship; and lastly, an RCT on the recruitment and supervision of teachers in Kenya

proved highly debatable, whereas they had been largely instrumental in elevating RCTs to the status of gold standard.

Following this preliminary research, we pressed on in two parallel directions. We took forward our work on rural microcredit in Morocco by conducting a replication. The results of this replication not only corroborated the hypothesis of a contradiction between RCTs in theory and RCTs in the field, but revealed new facets of this discrepancy (Bédécarrats et al. 2019a, 2019b). Expanded to a set of other RCTs on microcredit, this contradiction is the subject of one of the chapters in this book (Chapter 7). Keen to deliberate on the issue and prompt a scientific controversy, or at least discussion, we then launched this project to produce a co-authored book to throw open the question to other disciplines, voices, and opinions, including much more positive views of the method than ours. Some will argue that the debate is tiring and jaded (Dimova 2019; see also Ogden, Chapter 4). We believe, however, that is it vital, both scientifically and democratically speaking, for reasons detailed later in this book.

Bringing together some of the leading specialists in the field from a range of backgrounds and disciplines (economics, econometrics, mathematics, statistics, political economy, socioeconomics, anthropology, philosophy, global health, epidemiology and medicine, policy-making), this edited volume discusses the main weaknesses of RCTs in the field of development, but also some of their unexpected strengths. The book takes concrete examples to explain how RCTs work, what they can achieve, why they sometimes fail, how they can be improved, and why other methods are both useful and necessary. It reviews issues of method, epistemology, ethics, theory, and ideology. What stands it apart from other critical views is its emphasis (among others) on the implementation of RCTs *on the ground*, outside of their ideal laboratory conditions. This reveals some of their unsuspected uses and effects, their political uses and ends, but also their disruptive potential. The book explores the implicit worldview that many RCTs draw on and disseminate. It probes the gap between the method's narrow scope and its success worldwide. Yet it also proposes areas for improvement and alternative methods. Without disputing the contribution of RCTs to scientific knowledge, this book warns against their so-called superiority and the potential dangers of their misuse. It also argues that the best use for RCTs is not necessarily that which immediately springs to mind and which RCT proponents promote: understanding certain behaviour rather than evaluating interventions.

Although the principle of RCTs in science is over a century old—their use in international development is called the fourth wave (Jamison 2017)—their large-scale use in developing countries is unprecedented (Ravallion, Chapter 1). RCTs represent an indisputable advance for development economics. They offer a solution (among others) to the thorny question of attribution (how to isolate the effect of an

(Duflo, Dupas, and Kremer 2015), wherein Bold et al. (2013) have shown that scale-up by a national government-implemented policy produced none of the expected results.

intervention from all the changes that occurred at the same time). They place centre stage the issue of aid evaluation and the need for aid accountability. They lend new momentum to first-hand survey data collection by development economists. Last but not least, economic research in the past sidelined Southern countries due to their lack of quality data, especially longitudinal data. The spread of RCTs has elevated economic research on these countries to world-class level. The new wave of RCTs in development can also be interpreted as methodological progress initiated in the South and transferred to the North (Bédécarrats, Guérin, and Roubaud 2019).

Yet despite their limited scope of application (detailed below and throughout the book), RCTs are still held up by many as the evaluation *gold standard* against which all other approaches are to be gauged, and the award of the Sveriges Riksbank prize is likely to reinforce this supremacy. Presented by their disciples as a true Copernican revolution in development economics,[7] RCTs are often the only approach to be proclaimed "rigorous" and even "scientific" (see Ravallion, Chapter 1). Some media-celebrity RCT advocates are looking to take RCTs well beyond their methodological scope in a move to establish the full list of good and bad development policies (Labrousse, Chapter 8). The motive advanced for this upscaling ambition is to build up an ever-growing number of impact studies from which scalable lessons can be drawn. Clearly, though, there are a certain number of drawbacks to the proclaimed supremacy of RCTs in evaluation. These include disqualification and crowding out of alternative methods, ever-growing use of allocated resources, rent position, and the legitimization of a specific and narrow vision of "development" (what Lant Pritchett, Chapter 2, calls "kinky development"). They also include the disqualification of development projects and policies that do not adhere to the constraints demanded by the randomization protocols (Ravallion, Chapter 1; Garchitorena et al., Chapter 5; Patnaik, Interviews, this volume; see also Adams 2016).

We are obviously not the first to express criticism. Many voices have been raised.[8] James Heckman and Angus Deaton's critical voices (Deaton, Introduction,

[7] "Just as randomized evaluations revolutionized medicine in the 20th century, they have the potential to revolutionize social policy during the 21st," (Duflo, Glennerster, and Kremer 2004: 29).

[8] See for instance (Barrett and Carter 2010; Deaton 2010a; Deaton and Cartwright 2018; Harrison 2011; Heckman 1992; Pritchett and Sandefur 2015; Rodrik 2009). Several edited volumes have also contributed to this discussion. The first, a book edited by Jessica Cohen and William Easterly (2010), sparked the nascent controversy. The book contained just one chapter focusing specifically on the subject with a gripping, albeit brief, controversy between Banerjee, Rodrik, Mulathain, and Ravallion. The other chapters discussed mainly how to learn whether and which development policies work, but the question of RCTs ran implicitly throughout. The book by Tim Ogden (2017) is the most recent and RCTs are its central focus. It is structured in the form of 20 interviews with prominent players in the field. Fourteen of these players are active figures in the RCT movement and four others are more moderately involved in RCTs. There are just two critical voices (Angus Deaton and Lant Pritchett) who, although icons, make quite short contributions whose content is copiously reinterpreted and criticized by the other contributors. Thirdly, the book edited by Dawn Teele (2014) is more detailed and balanced. It makes a major contribution to our understanding of the subject, in particular with a comparison of RCTs conducted in the North and the South by political scientists and economists. Nevertheless, it still centres on methodological and epistemological considerations. Many contributions are repeats of now-dated articles published elsewhere in the 2000s, before the RCT

in this volume; Deaton, 2010a; Deaton and Cartwright, 2018; Heckman, 1992 and Chapter 12, this volume) carry particular weight, especially given that both have also been awarded the Sveriges Riksbank Prize in Economics (Deaton in 2015 and Heckman in 2000). This criticism is now more frequently acknowledged by RCT movement members (Ogden, Chapter 4), but there has been no actual scientific controversy over the issue. For want of a real controversy (the most eminent *randomistas* we invited declined to take part), this book creates a dialogue between approaches, disciplines, different intervention sectors, and ultimately different standpoints on the role and potential of RCTs.

Some of the book's authors consider that the RCT craze is "madness" (Pritchett, Chapter 2), that their superiority is essentially "a narrative" (Labrousse, Chapter 8), and that they are "ineffective as tools of organization accountability and learning," and are not strictly speaking evaluations (Picciotto, Chapter 9). Others consider that they have their place in the toolkit of evaluation methods, but that their self-styled superiority is "more a matter of faith than science," and that, in certain situations and for certain issues, observational studies are much more appropriate (Ravallion, Chapter 1). This is also shown by the sector analyses of healthcare (Garchitorena et al., Chapter 5), rural sanitation (Spears, Ban and Cumming, Chapter 6), microcredit (Bédécarrats, Guérin, and Roubaud, Chapter 7), and governance (Natarajan, Interviews, this volume).

A more optimistic view suggests that RCTs have taken on board the criticism and that, in their present version, they offer real answers to a large number of development questions (Ogden, Chapter 4). Another vantage point is that RCTs are useful not so much to "evaluate" as to "explore" behaviour using manipulations of price structures, contracts, teaching methods, and so on: researchers can make use of the disruption created by randomized protocols to observe *in situ* changes to interventions and behaviour, study their repercussions and draw operational conclusions from them (Morduch, Chapter 3).

Others call for them to be improved as much from an ethical point of view, which remains a blind spot for survey protocols in development economics (Abramowicz and Szafarz, Chapter 10), as from the point of view of causal explanation, whether with respect to making better use of priors (Vivalt, Chapter 11) or the phenomena of non-compliance as indicative of the preferences of targeted populations (Heckman, Chapter 12, this volume).

industry really took off. Ten years on, then, this co-authored volume brings the previous books up to date, drawing on the most recent literature and taking a broader view in terms of both disciplinary angles and issues. Finally, late 2019 at the time of finalizing our manuscript, *World Development* journal proposed a special issue on RCTs in development (to be published early 2020). Taking advantage of the attribution of the Sveriges Riksbank Prize in Economics to Barnerjee, Duflo, and Kremer, it gathers a bit more than 50 short notes (one or two pages) from a broad range of authors, Obviously, given the condensed format of the contributions, it cannot provide in depth analysis. However, apart from scaning a large spectrum of positions vis-à-vis RCTs (what did work and what did not), one of the special issue main interest is to propose avenues for future research.

The purpose of this introduction, which reflects solely the editors' point of view, is not to reconcile the authors and find a compromise, but to give readers a clearer picture of the issues involved in the debate. The first part details the epistemological, political, and ethical arguments behind the debate. The second part endeavours to define the development policies and projects that might lend themselves to the particularities of RCTs. The third part comes back to the idea of a scientific controversy, which we call for in earnest and which unfortunately has not yet taken place, looking into the reasons for this no-show. The conclusion proposes ways of improving RCTs and methodological alternatives.

0.1 The Arguments behind the Debate: Epistemological, Political, and Ethical

We will not go into all the criticisms made to RCTs here—they are already listed in different chapters (Ravallion, Chapter 1; Ogden, Chapter 4; see also Bédécarrats, Guérin, and Roubaud 2019). We think it more useful here to look over the epistemological, political, and ethical differences underlying—often implicitly—many of the disagreements surrounding RCTs.

Far from being purely technical debates, the debates surrounding RCTs make reference to different—and often hard to reconcile—concepts of knowledge and learning. Is social science research into human interactions perceived as scientism (Putnam 2009),[9] as the search for the ultimate, universal answer to a given problem, or as an ongoing learning process to find reasonable responses limited in time and space, mindful of the diversity of knowledge, including the knowledge of the development target populations? Do we see figures, statistical and econometric methods applied to social sciences solely as instruments and techniques, as the fruit of linear scientific progress? Or do we consider them also as a social and political construct built by somewhat arbitrary conventions, inextricably linked with a certain conception of state and public policies, the market, power, and collective action (Desrosières 2013b), which fashion in part the world they seek to represent, understand, and advise (MacKenzie, Muniesa, and Siu 2007)? This second meaning of knowledge does not deny scientific evidence, but advocates its embeddedness in particular social and political contexts. And it clearly differentiates scientific knowledge from policy decision-making, which implies referring to values in order to choose between different options and assess their social, economic, and political consequences (Drèze 2018a).

[9] By scientism, we are referring to the idea that experimental science is the only reliable source of knowledge on the world and that it is the best means by which to organize humanity to solve all its more pressing problems. Experimentation allegedly does without the need for metaphysical, philosophical, ethical, and aesthetic reasoning.

The opposing views surrounding RCTs are also based on different notions of development, poverty and, more broadly, politics, seen as a conception of the world in which we live and which we endeavour to attain. Is the world an aggregate of individuals seeking to be independent or is it a complex system made up of dialectics, multiple interactions, retroactions and systemic effects between social beings who are interdependent and wish to remain so? Should we see the "causes of poverty as a lack or want of relevant variables or as an active process of impoverishment or perpetuation of poverty" (Shaffer 2015: 154)? A "want-based" understanding of the causation of poverty calls for policies of "difference-making" wants (to cope with deficits in health, education, nutrition, water/sanitation, credit, and so forth); and understanding the impacts of such policies requires a counterfactual to be able to isolate the difference and attribute the impact to the policy in question. By contrast, a conception of the causation of poverty in terms of processes and social relations calls for macroeconomic and structural policies (exchange rate, capital control policies, social protection measures, and so forth); and understanding the impact of these measures requires a "mechanism-based approach" that explores the diversity and complexity of the causal processes that generate the impact (Shaffer 2015).

Finally, these divergent visions find expression in divergent versions of the economists' role. Is their role to "fix" the world and concentrate on the practical details of policy implementation (Duflo 2017), like a plumber or engineer repairing cracked pipes? Or should economists keep a critical distance from the workings of the present system, even going so far as to radically challenge it?

These different epistemological positions (in the form of a continuum more than a binary opposition) permeate the debates on RCTs and can be seen in a string of opposites running through the chapters of this book: macro versus micro, public goods versus private goods, horizontal versus vertical health interventions, public action versus social marketing, structure versus behaviour, attribution versus processes, and so on (Ogden, Chapter 4; Labrousse, Chapter 8).

0.1.1 The Epistemology of RCTs in the Field of Development

In theory, *randomistas* see experimentation precisely as an antidote to preconceived ideas (see also Rodrik 2009). This pragmatism may well give the impression of being a rejection of scientism. Yet laying claim to the method's superiority clearly reflects a scientistic concept of science (Picciotto, Chapter 9). This scientism can be seen at work in two ways. First of all, the *randomistas* purport to provide universal answers for a large number of development interventions. In response to the question of contextual particularities, some *randomistas* like Esther Duflo argue that they should be considered as "global public goods" and an international body established to scale them up (Savedoff et al. 2006; Glennerster 2012). This body

would then build a universal database and act as a "clearing house," providing answers as to "what works and doesn't work" in development (Banerjee and He 2008; Duflo and Kremer 2005). Yet this hegemonic plan (Bédécarrats, Guérin, and Roubaud 2019) does not solve the question of heterogeneity, whether of intervention practices or contexts (see, in particular, Spears, Ban, and Cumming Chapter 6).

Secondly, this scientism is seen at work in overconfidence in the technique, with something of an obsession with the *theoretical* protocol, supposed to guarantee sample balance and therefore settle the attribution question. The *implementation* of the protocol on the ground is secondary. As with all research—particularly RCTs considering the budgets concerned, the size of the samples, the constraints for comparison between control and treatment groups, and the risks of contamination—the implementation of the protocols necessarily deviates from what is planned in theory and calls for tweaking, accommodations, and compromise.[10] In many cases, the collection of RCT data violates the assumptions of the statistical theorems used for inference. NGOs and governments working in development know only too well that interventions in the field never go according to plan (Mosse 2004; Olivier de Sardan 1995). Why should experiments be any different? As shown by the different chapters in this book, deviations between protocol and implementation can be observed all the way down the knowledge production line:

- In sample building with, here, three types of difficulties. The first difficulty is multiple biases between treatment and control groups (Ravallion, Chapter 1). This results in a focus on highly specific populations, although this particularity is not made clear by the randomistas (see, for example, Bédécarrats et al. (2019a) and Wydick (2016) on microcredit; see also Barrett and Carter (2014: 75), Moatti, Interviews, this volume). The second difficulty is insufficient take-up and consequently an insufficient difference in exposure to the intervention. This weakens the ability to draw conclusions due to a lack of statistical power, a problem that would require unrealistic sample sizes and therefore unrealistic budgets to resolve (McKenzie 2012; Spears, Ban, and Cumming, Chapter 6). Insufficient take-up can also cause the intervention to be artificially transformed (see the following point). Lastly, the "virginity" of the control zones, an often necessary condition for comparison, proves particularly complex and raises ethical and feasibility problems (Bédécarrats, Guérin, and Roubaud 2019).

[10] Our replication experiment shows the difficulty some *randomistas* have acknowledging the practical difficulties of conducting an ideal RCT, the like of which does not actually exist (Bédécarrats, Guérin, and Roubaud, Chapter 7).

- In the type of intervention, whose implementation may turn out to be very different to the "real world," as shown by Garchitorena et al. (Chapter 5) in health, or which may even be artificially transformed to encourage more take-up (Bédécarrats, Guérin, and Roubaud, Chapter 7).
- In data collection, since the priority placed on econometric considerations can get in the way of statistical considerations. Statistics is not only the science of numbers: it is first and foremost a science of data collection, which requires multiple techniques to guarantee the collection of quality data (Bédécarrats, Guérin, and Roubaud, Chapter 7).
- In the interpretation of the results which, far from being restricted to a comparison of averages, as claimed by the randomistas, actually implies a range of implicit hypotheses and an art of rhetoric, whose persuasive power is particularly manifest (Labrousse, Chapter 8).

All in all, method implementation constraints can force researchers to concentrate on midpoint indicators, short timeframes, and specific populations or geographic areas and, in so doing, to restrict themselves to a very narrow set of questions or produce unusable results (see Chapters 5, 6, and 7, this volume, on different sectors give numerous examples of this; see also the case of public health (Moatti, Interviews, this volume) and governance (Natarajan, Interviews, this volume). The disproportionate importance placed on the theoretical purity of the protocols and demonstration of causality at the expense of protocol *feasibility* and data *quality* is a major (albeit often implicit) sticking point in disagreements over the hierarchy of methods.

From our point of view, giving precedence to the method over the research questions is tantamount to "hunting for the lost keys under the streetlight." In a way, and to paraphrase the title of a book on development aid (Naudet 1999), it is like finding problems (projects to evaluate) to the solution (RCTs).

0.1.2 RCTs and "Development"

As suggested by Lant Pritchett (Chapter 2), the success of RCTs is merely the symptom of a more serious disease: the abandonment by part of the international aid community of large-scale transformative development policies (national, international, and even regional), including seeking to transform the socioeconomic systems.[11] Reviewing transformations in the field of aid is therefore useful to better understand the attraction of RCTs and their scope of application. The

[11] The prize jury, in its press release, acknowledges this: "This year's Laureates have introduced a new approach to obtaining reliable answers about the best ways to fight global poverty. In brief, it involves dividing this issue into smaller, more manageable, questions."

contrast between the narrow scope of RCTs and their scientific, media, and political success is down to both supply and demand. On the supply side, we have shown elsewhere that the *randomistas* have produced an entirely new scientific business model, of which J-Pal is the most emblematic and accomplished example, and which combines 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, Guérin, and Roubaud 2019). As effective as these strategies may be, they nevertheless assume that there is a *demand*. Some methods, theories, and technologies succeed, not because of their scientific superiority, but because they manage to "sustainably galvanize and rally players and interests prepared to produce and use [the technologies in question]" (Callon 2006a: 155).

RCTs benefit here from a particularly RCT-friendly environment, which they nurture in return. They most probably would not have had the same success in a different age. The academic climate first of all, especially in economics, is conducive to the rise of RCTs: demise of the heterodox schools concentrating on social structures and domination processes, search for the micro-foundations of macroeconomics, and primacy of quantification and economics in the social sciences. The joint rise of behavioural and experimental economics, crowned by the 2002 award of the Sveriges Riksbank Prize in Economics to psychologist Daniel Kahneman and economist Vernon Smith, respective experts in the two fields, and then to economist Richard Thaler in 2017, shows just how far the discipline has come. RCTs draw extensively on the precepts of behavioural economics and are actually the vehicle that channelled behavioural economics into development economics to the extent that it now occupies a dominant position in the discipline (Fine et al. 2016).

It is also from transmutations in the aid field that demand has emerged for RCTs. With the end of the Cold War, the political sphere started to ease its grip on official development assistance (ODA). Cold War technical and financial cooperation was often merely another pawn in bloc rivalry. As the Berlin Wall fell, so too did cooperation's subordination to realpolitik. In the new post-modernist world, ODA promoters have found themselves under the spotlight as the aid crisis, MDGs, and *New Public Management* have summoned them to the stand to prove their utility (Naudet 2006).

The new credo focuses development policy on poverty reduction and promotes results-based management. These guidelines were formulated in the 2005 Paris Declaration on Aid Effectiveness and thereafter systematically reiterated by the major international conferences on official development assistance in Accra in 2008, Busan in 2011, and Addis Ababa in 2015. The rise of the evidence-based policy paradigm, which consists of basing all public decisions

on scientific evidence, has given scientists new credibility in these political arenas. RCTs in principle meet all the conditions required by this game change: agnostic empiricism, apparent simplicity (simple comparison of averages), elegant use of mathematical theory (guarantee of scientificity), and focus on the poor (compassionate mobilization and moral commitment; Labrousse, Chapter 8). Their (apparent) simplicity makes them easy for policy-makers to understand, lending them appeal as a vehicle for informing public decision-making. The evaluation of the *Progresa* programme in Mexico formed a prototype for this method and a textbook example of its performance capabilities (Bédécarrats, Guérin, and Roubaud 2019).[12]

The aid crisis is also a crisis of *official* development assistance. As ODA funding efforts lose speed, private investment, and international remittances are taking up the slack (IFC 2017). Governments are now merely one body among others in a "coalition of players" that includes businesses, NGOs, and, more broadly, "civil society," foundations and research institutes. Foundations taking up the philanthrocapitalism of the industrial period are playing a growing role, mainly in the health sector, but also in technological innovation, now cross-cutting most, if not all development sectors (see also de Souza Leão and Eyal 2020). These new players and funders are changing the aid *tools*. Not only does the withdrawal of the State as planner and developer lead to "thinking small" (Cohen and Easterly 2010), but when combined with the resurgence of philanthropy, it paves the way for development that juxtaposes privatization (of interventions and players), marketization (of the goods and services delivered), and also compassion.

By setting up the poor as barefoot entrepreneurs, microcredit with its promise of a double bottom line—poverty reduction with profitability or at least financial sustainability—was a pioneer in marketization. This marketization subsequently expanded under the name of BoP (bottom of the pyramid) in a low-cost repeat of trickle-down theory (with the idea that consumption by the poor will eventually form a factor for growth and redistribution (Elyachar 2012)).

This economic reason combines with a "humanitarian" reason (Fassin 2010). A moral duty to act is emerging in the face of public infrastructures seen as moribund, derelict, or utopian and the resulting suffering and needs they create. Driven by a sense of compassion and urgency, financiers and practitioners—but also researchers—are joining forces to design and test an entire array of micro-scale interventions: these "humanitarian goods," to use the expression coined by Redfield (2012), try as best they can to solve, ad hoc and temporarily, what are considered to be the most urgent and crying needs. These humanitarian

[12] It is, however, enlightening to note that this programme was a powerful tool for social and political control, consumed by nepotism and corruption (Crucifix and Morvant-Roux 2018; Kidd 2019). Moreover, the blind spots in its experimental evaluation, especially in terms of internal validity (Faulkner 2014), were precisely the arguments used by the new Mexican government to announce its withdrawal in early 2019 (Encisco 2019).

goods aim to mitigate government failings, and they embrace this dual compassionate and economic line, even though the economic strand does not rule out redistributive measures (see **p. 18**).

In this new configuration, and although public financing of large infrastructure continues to account for a large share of international aid, governments' decision-making and planning powers are gradually seeping away towards vertical funds,[13] foundations[14], private companies,[15] and new financial mechanisms such as social impact bonds. The foundations, a fast-growing emerging player, are set to play an increasingly important role (see also Pritchett, Chapter 2). Just as the Ford Foundation supported the rise of experiments in the United States in the 1960s, so too are numerous foundations today playing a driving role in the expansion of RCTs in development (starting with the establishment of J-Pal; Jatteau 2016: 230). The very principle of social impact bonds, in which repayment to investors is conditional upon specified social outcomes being achieved, favours a similar trend. Lastly, in this development privatization process (privatization of interventions and players alike), NGOs occupy a choice position as implementing partners.

Far from the reforming and sometimes idealistic aims of previous generations of development players, private, market and humanitarian goods have the merit of being realistic and concrete and offering a pragmatic solution for needs seen as urgent. Their implementation is not above criticism—probably the most well-known are the debates on therapeutic food as an unfair trade practice impacting on local agricultural systems. Yet from the point of view of their purpose—to solve a temporary, individual problem—they work (Redfield 2012). Now, as a number of chapters point out, and we will come back to this later, it is precisely these types of goods, due to their individual targeting and short-term nature, that lend themselves the best to the constraints of randomized trials. Likewise, NGOs remain the choice implementing partners for *randomistas*, because they are more flexible, less bureaucratic, more open to innovation and more reliable than governments (Webber and Prouse 2018; Cohen and Easterly 2010). *Randomistas* express a will to work more with governments (Banerjee 2013), but are finding it hard to deliver on this will (Pritchett, Chapter 2). In India, a privileged field of study for RCTs, the testimonies of a senior Indian official (Natarajan) and a prior principal economic advisor to the Government of India (Patnaik) in the Interviews (this volume) suggest that the impact of RCTs on policy-making is not

[13] Such as the Global Fund and the GAVI Alliance (Global Alliance for Vaccines and Immunization) in health.

[14] The Bill & Melinda Gates foundation remains the leader in many health subsectors, but also in everything involving new technologies. Banking foundations such as Citi and Mastercard are high-profile players in financial inclusion.

[15] Such as Nutriset for therapeutic food to treat malnutrition and Vestergaard Frandsen for water filters, tsetse fly screens, and insecticidal bednets.

only negligible but counterproductive, since RCTs distract from real issues and weaken the economics profession.

The transformation of the development field came well before RCTs, and it would be undue to say that they were responsible for it (Morduch, Chapter 3), even though the crowding-out effects are to be taken seriously (see Section 0.3.2). Yet should these changes be unreservedly condemned by an indictment of the abandonment of any true prospect for reform, the unsustainability of individual ad hoc interventions, and the illegitimacy of private players who are not democratically accountable? Or should we learn to live with them rationally, considering that even though the pipes were poorly designed to begin with or are at breaking point, to use the plumber/engineer metaphor, it is still worth repairing the leaks? The answer to this (rarely spelled out) question explains many of the disagreements surrounding RCTs, as well as the different positions found in this book.

0.1.3 Ethics and RCTs

The ethical issue is a recurring one with RCTs, not only in the development field but in general (especially in medicine). Although everyone agrees on the need to tackle this question head on, at least in principle, these caveats have not yet seen action (Abramowicz and Szafarz, Chapter 10; and also Ravallion, Chapter 1; Ogden, Chapter 4; Bédécarrats, Guérin, and Roubaud, Chapter 7; Picciotto, Chapter 9; Patnaik, Interviews, this volume). Among the *randomistas*, this acknowledgement remains marginal,[16] as if faith in the scientific advances that RCTs can bring—and their automatic policy and welfare improvement repercussions—were sufficient to exempt researchers from ethical consideration. Whereas all research entails ethical issues, RCTs are more concerned than observational studies by reason of their very principle (Teele 2014), since they typically feature a form of manipulation of the research environment (they "twist the lion's tail," to quote the expression used by Deaton and Cartwright (2018: 18)).

Neither are critical analyses free of this neglect of ethical considerations, since they often merely mention the issue with barely any details. The chapter by Michel Abramowicz and Ariane Szafarz (Chapter 10) is an exception in that it probes the implications of the principle of *equipoise*, i.e. the ethical requirement for an experiment involving human subjects to display "a state of genuine uncertainty on the part of the clinical investigator regarding the comparative therapeutic merits

[16] For example, none of the 22 pages on "Concerns about experiments" by Banerjee and Duflo (2014) addresses the ethical issue, except to say in response to randomization not being a fair way to allocate the programme (seen as a methodological problem, but not an ethical issue) that "implementers may find the easiest way to present it to the community is to say that the expansion of the programme is planned for the control areas in the future" (p. 101).

of each arm in a trial" (Freedman 1987: 141, quoted by Abramowicz and Szafarz). The authors ask why economics experimenters are virtually systematically ignorant of this principle, when it is an essential pillar in medical science. They provide a series of pointers to address the question. Ravallion also addresses this subject (Chapter 1), insisting on the importance of properly assessing the risks and information already available, and showing that the principle of equipoise takes different forms depending on the different cases and types of randomization (inevitable treatment rationing, conditional randomization, and equivalence trials). He also discusses the *adaptive experiment* proposition put forward by Narita (2018) to establish a Pareto balance between the possible positive and negative effects on participants based on available knowledge.

This virtual denial of ethical considerations by *randomistas* is all the more questionable in that various standards of best practices do exist, as much for medical RCTs as for most of the social science RCTs conducted in the North. The ethical principles designed to govern randomized trials on human subjects have been codified into recognized standards, in particular the Declaration of Helsinki in 1967 (WMA General Assembly 2014 (9th edition)); the Belmont Report in 1974[17] and the International Ethical Guidelines of Council for International Organizations and Medical Sciences (2002). These standards prescribe clear principles: informed consent, the do no harm principle, provision of specifically considered protection for vulnerable populations, risk analysis, and responsive monitoring, to name but a few.

These fundamentals are rarely respected in the development field (Abramowicz and Szafarz, Chapter 10). Like Barrett and Carter (2014), we detail four examples of RCTs that illustrate the ill effects of this ethical negligence. The first example was designed to demonstrate the mechanisms of corruption in the case of obtaining a driver's licence in India (Bertrand et al. 2010). One of the arms of the treatment was to offer a bonus to candidates for obtaining a licence. Barrett and Carter show that this RCT violated the ethical code of "do no harm" (they even speak of "irresponsible research design") in two ways: not only did the treatment encourage corruption, but it also imperilled the lives of others by putting potentially reckless drivers on the roads, since the experiment showed that the treated group took fewer driving lessons. The second example concerns an RCT set up in Kenya to test the Rockefeller Effect (which states that too many resources do more harm than good) by means of a project providing assistance to groups of women (Gugerty and Kremer 2008). The project's effects proved to be negative (the poorer women were excluded from the positions of power), confirming the Rockefeller hypothesis. The problem is that the RCT harmed the experiment's subjects, when this harm might have been predicted, at least as a possibility, and

[17] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979).

the women should have been informed of that possibility for them to decide whether or not to take part (principle of informed consent). The third case is an RCT on secondary school pupils in the Dominican Republic to test whether information concerning higher labour market returns to education than would normally have been expected by the pupils could prompt them to stay longer in education (Jensen 2010). The ethical problem here is that the information given the secondary school pupils (estimated from observational data)—being both overestimated (given the endogeneity biases) and calculated on average without taking into account the characteristics of the pupils and schools—is likely to have led some of the pupils, probably the poorest, to "overinvest" in education on the basis of the expected return. And that is not to mention the effect of the increase in the supply of graduates likely to depress future returns (general equilibrium effect). Last but not least, the treatment in the fourth example consisted of granting credit to individuals rejected by a microcredit provider, since its scoring model predicted a high probability of default on payment (Karlan and Zinman 2009).[18] Quite aside from the fact that this strategy placed the treated group at risk of being incapable of reimbursing the loans (with the associated penalties) and in a potential situation of over-indebtedness, not having informed them of such is in breach of the principle of informed consent. This is a difficult dilemma to resolve since, if they had been informed of the risk, their behaviour would probably have changed and hence undermined the internal validity of the RCT. None of these flaws prevented these four RCTs from being published in leading academic journals, and this also raises questions about the role of economic journals in failing to respect ethical standards (Abramowicz and Szafarz, Chapter 10).

Other examples are mentioned in the book (Ogden, Chapter 4; Abramowicz and Szafarz, Chapter 10; Patnaik, Interviews, this volume). The proliferation of RCTs, especially by less visible and hence even less ethically controlled institutions, could end up undermining the basic principles. Illustrating this point is the case of an RCT in progress. This donor-commissioned RCT was set up to test how information affects migratory behaviour in rural Mali. Participants were shown a short film chosen from among four randomly allocated films illustrating different outcomes of migration and non-migration (to Europe): successful migration; suffering, ill-treatment, and ultimately a failed migration attempt; successful non-migration; and a comedy having nothing to do with migration serving as a placebo. Quite aside from the problems of informing participants of the implications of such a test and obtaining their informed consent, none of the treatment arms can be deemed beneficial to the participants (violation of the beneficence principle). Individuals' preferences can simply be changed based on

[18] A similar approach consisting of including subjects initially judged insolvent is also included in Augsburg et al. (2015) and discussed by Bédécarrats Guérin, and Roubaud, Chapter 7.

what the RCT's commissioners consider to be good for them (or for themselves). Neither is the "do no harm" principle respected as, following the films, some participants might decide to migrate and die in the Mediterranean or be tortured in Libyan jails. Lastly, the commissioner's political motive seems obvious (to curb African migration to Europe). It seems that this RCT was designed without serious ethical consideration.

Considering the multitude of examples, it would appear that the creation of Institutional Review Boards in many academic institutions has done nothing, or at least insufficiently, to remedy the observed ethical lacunas (Barrett and Carter 2020). Two mutually reinforcing reasons can be put forward for this. The first is the difficulty of simultaneously guaranteeing the protection of the experiment's subjects and the internal validity of the protocol. The second is the *randomistas'* flawed understanding of and manifest lack of interest in the subject. When faced with what could be called an ethical dilemma, they all too often come down on the side of the methodological imperative. Yet ethical safeguards are all the more necessary in the Southern countries. Firstly, not informing participants (informed consent principle), if not deliberately misinforming human subjects to ensure a clean identification strategy, is at odds with the principle of ownership promoted by the development policies. Secondly, participants are generally vulnerable individuals, both economically (poor) and politically (voiceless), on whom it is easier to impose the trial, if not deliberately mislead. This asymmetry is especially strong in that the surveys are more often than not tantamount to lifesize laboratory games supervised by young students and research assistants from Northern universities. We also need to look into the choice of these populations, especially when testing a behavioural hypothesis or a theory put forward by certain RCT proponents (Banerjee and Duflo 2011; see Morduch, Chapter 3). Save advancing that the poor in Southern countries have specific rationality, the arguments of lower cost and less capacity to refuse to take part (a recurring problem with RCTs in Northern countries) due to a lack of knowledge of their rights and lopsided balances of power (including with respect to the experimenters) appear to be credible explanations (Patnaik, Interviews, this volume; see also Teele 2014), as has already been observed in the "offshoring" of medical clinical trials (Petryna 2007). Without going so far as to call for a "moratorium on experimentation" in the South (Hoffmann 2020), the issue should be at least addressed in priority.

The *randomistas*' ethical argument is the long-term improvement of the well-being of populations by means of scientific progress made possible by RCTs. Yet this is an assumption that is far from proven (Ravallion, Chapter 1). All in all, then, in addition to the unassailable faith in the theory of the technique at the expense of its feasibility (as seen in Section 1.1), it seems that all too often, a hardly acceptable hierarchy of values prioritizes scientific findings over the well-being of the populations.

0.2 What Is the Scope of Application for RCTs?

After closely examining the many limitations of RCTs, in terms of both internal and external validity, Deaton and Cartwright (2018) suggest that RCTs nonetheless remain valid in two areas: (1) to test a theory, and (2) for a specific evaluation in a given context of a particular project or policy, provided that the potential internal validity problems have been solved and with the caveat that the explanation of the results obtained is often inadequate. The chapters in this book confirm and expand on this analysis. Randomized evaluations are only possible for a highly restricted field of interventions, more often than not concerning private, market, and humanitarian goods. RCTs can also be used to test economic theory regarding behavioural responses to interventions, challenging certain preconceived ideas. Ultimately, however, they answer neither the question of *impact*, such as it has long been defined in the development aid field, nor the question of the *explanation* for the measured effects.

0.2.1 Private, Market and Humanitarian Goods

The conditions required by the randomized methods' protocols restrict them to a narrow spectrum that Bernard, Delarue, and Naudet (2012) call "tunnel-type" programmes. These programmes are typified by short-term impacts, clearly identified, easily measurable inputs and outputs, and unidirectional (A causes B) linear causal links, and are not subject to the risks of low uptake by targeted populations. They tie in with the suggestions made by Woolcock (2013) that projects subjected to randomization need to exhibit "low causal density," require low implementation capability and feature predictable outcomes.

This type of method is therefore applicable only to simple or local short-term interventions targeting individuals. In concrete terms, these micro-interventions concern essentially private goods and services, i.e. rival and excludable (see Ravallion, Chapter 1; Pritchett, Chapter 2 and Picciotto, Chapter 9).

In health, they concern actions to prevent and treat individual diseases. They also come in the form of water filters, mosquito nets, training, and bonus systems for health professionals, free consultations, medical advice by text message, and micro-insurance. However, RCTs do not answer the question of the management of the health systems, which are necessarily complex and systemic, involving skilled, motivated manpower, an infrastructure, the provision of medicines, etc. (Garchitorena et al., Chapter 5). In sanitation, these micro-interventions concern the distribution, construction, and use of latrines. Here again, RCTs do not answer the question of the management of human waste flows using which type of sanitation or cleaning network, which type of infrastructure, and which type of regulation (Spears, Ban, and Cumming, Chapter 6). In poverty reduction, these

micro-interventions are microcredit, savings, entrepreneurship training, and financial education services. Once again, RCTs do not answer the question of regional or sectoral wealth creation processes (Bédécarrats, Guérin, and Roubaud, Chapter 7) or the broader question of access to basic services (Pritchett, Chapter 2). In governance of public administrations and institutions, these micro-interventions are random inspections, financial incentives, independent third-party audits, and call-centres and telephone feedback. RCTs do not answer the question of weak state capacity, centralized bureaucracies marked by low trust, scarce resources, over-burdened bureaucrats, and challenging work environments (Natarajan, Interviews, this volume).

Contrary to certain critical analyses (see, for example, Berndt 2015), RCT conclusions do not necessarily advocate the marketization of the private goods (which equates RCTs more with the above-mentioned humanitarian camp). In the case of highly price-elastic insecticidal bednets and deworming treatment, RCTs have put the case precisely for their free distribution, considered to be more effective than billing and hence challenging popular belief in the health field. In the case of microcredit, RCTs have concluded that the poverty reduction impact remains marginal and that poverty reduction therefore calls for other types of intervention (Banerjee, Karlan, and Zinman 2015). Again in the case of microcredit, RCTs have shown that the poor are sensitive to interest rates, here too toppling the widely held idea that access is more important than cost, a popular belief among microfinance organizations and their financers held up to legitimize high interest rates (Morduch, Chapter 3).

Although these findings can be useful, the subjects addressed remain limited compared with the host of development, poverty, and inequalities issues. The conditions required to implement RCTs therefore rule out a huge number of development policies involving combinations of socioeconomic mechanisms and feedback loops (emulation effects, recipient learning effects, programme quality improvement effects, general equilibrium effects, etc.). This is precisely the case with public goods (Ravallion, Chapter 1). Where interventions involve infrastructures and regulatory systems, experimental manipulation is impossible (Spears, Ban, and Cumming, Chapter 6).

In the terms of reference for a study commissioned on the subject, a group of DFID managers estimated that less than 5 percent of development interventions are suitable for RCTs (DFID 2012). Although this figure is not to be taken literally, there is no doubt that experimental methods are not suitable to evaluate the impacts of the vast majority of development policies. In their more formalized paper, Pritchett and Sandefur (2013b) come to a similar conclusion.[19] In this volume, Garchitorena et al. (Chapter 5) point out that 97 percent of funding for

[19] "The scope of application of the 'planning with rigorous evidence' approach to development is vanishingly small" (Pritchett and Sandefur 2013b: 1).

health research worldwide is earmarked for the development of new technologies (mainly of the pharmaceutical variety), and that only the 3 percent left over goes into research on *implementation*, albeit essential to understanding and improving health system dysfunctions.

0.2.2 Evaluate Impact or Test Behaviour?

As suggested by Jonathan Morduch in this book (Chapter 3), RCTs actually pursue two aims: to measure impact, and to explore "the nature of economic contracts, behaviors, and institutions." He goes on to submit that it is ultimately this second less-debated type of "exploratory RCT" that is the most promising, representing a real gain over other methods and therefore greater potential in terms of expanding knowledge.

This second type of RCT shifts the focus from measuring the impact of interventions representative of public action or development aid to testing different modes of a given intervention and measuring the outcomes in terms of intervention take-up. This type of RCT, says Morduch, is a source of information, if not "provocation," in challenging certain misconceptions in development economics (such as the above-mentioned low price elasticity of demand for microcredit) and testing innovations and how behaviour reacts to those innovations. For example, it can test different crop insurance selling timeframes for a better understanding of the constraints of time and liquidity; or test the role of information and assistance in the use of mobile telephones by the ultra-poor for a better understanding of intra-household sharing mechanisms.

These purposes are useful and laudable (provided the ethical and internal validity criteria are met and the conclusions are valid), but the question could be asked as to why the *randomistas* persist in talking about impact when a large number of RCTs are actually more "exploratory" in nature and compare different modes of one and the same intervention, often merely measuring the take-up differentials. The sector analysis of sanitation comes to a similar conclusion: RCTs appear to be more suited to analyzing behavioural changes than measuring impact per se (Spears, Ban and Cumming, Chapter 6).

In fact, the question of impact often remains unanswered. Since 1992, most development aid sector players have relied on five criteria defined by the OECD Development Assistance Committee (2002), among which is found an impact criterion: "Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended." Yet RCTs can only evaluate the *short-term impact of short causal chains*: this is not then strictly speaking an impact as defined above (see also Picciotto, Chapter 9). Taking the example of insecticide-treated bednets, often seen as the jewel in the crown of RCTs (Ogden, Chapter 4), the question usually asked by

RCTs concerns *take-up* rather than impact, since insecticide-treated bednets are considered to be essentially "good." Yet their medium- and long-term effects are controversial due to genetic adaptation by mosquitoes and the destruction of local production systems (Beisel 2015). Omitting long-term and collateral effects is just as problematic in the microcredit sector (Bédécarrats, Guérin, and Roubaud, Chapter 7).

This type of RCT is ultimately redolent of the notion of "social marketing," a term very much in vogue in development circles, which quite naturally complements the above-described circulation of private, market and humanitarian goods and behaviourist trend. Social marketing is the application of commercial marketing tools and principles to the design, implementation, and evaluation of behaviour change programs in pursuit of individual benefits and the public interest (French et al. 2010). Modelled on behavioural science, social marketing techniques include nudges,[20] but also more classic marketing methods (packaging, price, identification of the most suitable distribution channels and places, etc.). Social marketing originated in the 1970s in the health and social fields, including in the South, and in areas such as reproductive health, AIDS prevention, rehydration therapy for diarrhoea, and sanitation), before expanding to target behavioural change in a large number of sectors (environment, agriculture, education, financial management, consumption, etc.).

0.2.3 Measuring versus Explaining

RCTs might be able to measure and test some intervention impacts and aspects, but they cannot analyze either their *mechanisms* or their underlying *processes*. In a "want-based" analysis of the causation of poverty, as found in randomized approaches, the question of processes and mechanisms is set aside (Shaffer 2015). Overcoming this limitation of the probabilistic theory of causality would call for a "causal model" (Cartwright 2010), a coherent theory of change (Woolcock 2013), a structural approach (Acemoglu 2010) and evaluation of the intervention in context (Ravallion 2009a, Chapter 1; Pritchett and Sandefur 2015).

In the face of this criticism, *randomistas* are now grounding their results in explicit theories of change (Ogden, Chapter 4), based largely on behavioural economics. Behavioural economics is useful to disentangle the complexity of the psychological and cognitive processes, individuals' internal struggles, and the multitude of their "mental accounting" practices (Thaler 2015), and to explore and test how behaviour reacts to such or such an intervention (see Section 0.2.2). However, behavioural economics cannot capture the complexity of atypical, unexpected and "suboptimal" behaviour, wherein this latter term moreover

[20] A nudge refers in behavioural economics to a small and cost-effective device that does not imply a formal obligation or prohibition intended to influence people's behaviour in a predictable way.

presumes what is a debatable normative dimension. Two levels need to be differentiated here: the level of individual behaviour, which does not always fit in the behavioural boxes (Servet 2018; Servet and Tinel 2020), and the level of the interventions, which rarely go according to plan.

With respect to individual behaviour, people are social, plural beings who cannot be reduced to mere target populations. People's agency is not limited to their refusal or take-up. Neither is it limited to their cognitive or social "biases." Local rationalities and motivations are constructed; they develop from and reflect social and political norms and realities. They tie in with pre-existing forms of interdependences, balances of power, and social and political structures, but also with desires and aspirations. The fact that the social is so unpredictable does not mean that it should be seen solely as an obstacle and a constraint, that it should be eliminated by dint of nudges. Local populations sometimes have good reasons to act the way they do, especially when the global environment does not change. People have their own conceptions and representations of the world (and their own theories of change) and their own knowledge and know-how regarding care, illness and well-being, cleanliness and dirtiness, finance, poverty and wealth, and so on. Although some of these representations are sources of discrimination, the fact remains that they shape behaviour. Some of these representations also reflect specific worldviews, which are not necessarily less "optimal" than the researchers' own (see the example of microcredit with Bédécarrats, Guérin, and Roubaud, Chapter 7).

Interventions are also complex, combining multiple levels and players. Local realities shape, frame, constrain, and influence the intervention (Mosse 2004; Olivier de Sardan 1995). Such is the case with the three sectors represented in this book. In global health, for example, "One of the most important questions […] is why known technologies – those that are proven to work in certain settings – systematically fail to reach the people for whom they are intended" (Garchitorena et al., Chapter 5). Answering these questions necessarily calls for a focus on the workings of "systems": local health systems, organizations' systems, the particularity of interactions between "target" populations and healthcare providers, etc. Likewise, sanitation and microcredit do not mean the same thing to different people and cover countless realities, methods, and forms of implementation. This diversity narrowly restricts the potential of RCTs in their generalization endeavours (Spears, Ban, and Cumming, Chapter 6; Bédécarrats, Roubaud, and Guérin, Chapter 7). And this complexity and diversity are probably not limited to these three sectors.

0.3 Why Is a Scientific Controversy Needed and Why Has It Not Taken Place?

As we have seen, far from being a unanimously accepted gold standard, RCTs are a subject of debate and much criticism. This should have prompted a scientific controversy. But there has not (yet) been any such controversy, when it is vital for

scientific progress and democratic debate. What has gone wrong? Without claiming to be exhaustive on the subject, and in view of its importance, we felt it useful to outline a few avenues for analysis borrowing from science studies.

Bear in mind, first of all, that contrary to a naïve view of science, scientific progress is not always a rational, linear process wherein the most effective and useful methods and findings systematically prevail over the others and a consensus surrounds certified knowledge. Scientific knowledge is also a historical, social, and political product forged by advances and setbacks, cycles, debates, and disagreements, which sometimes turn into *controversies*, defined as differences between two parties brought before and debated on a public stage.

A scientific controversy is therefore not to be understood in a negative light, as the symptom of errors of reasoning (where the "true" will ultimately prevail over the "false") or of untoward interference by politics or interests other than the advancement of knowledge (an area supposed to be free of all subjectivity). The controversy is inherent in the collective production of knowledge. It is often the controversy that enables the emergence of key scientific progress. All scientific fields are marked by major controversies that are sometimes violent (water memory, GMOs, the "Eldorado scandal," and gravitational waves), but are also sometimes nipped in the bud (Callon 2006a).

A controversy may be defined as a difference between two conflicting positions taking as witness an audience made up of scientific peers or a broader public (Lemieux 2007). The conflicting positions taken are sometimes virulent, but participants are bound to respect the conventions of the academic world such as the principle of equality between participants, the importance of logical reasoning, control of aggression, and respect for the principle of the dignity of the protagonists. However, these conventions remain vague, and accusing an opponent of abusing a dominant position or overstepping the limits of civility is often a way of shifting the balance of power or disqualifying the rival.

As in many areas of sociology, schools differ as to how to approach controversies depending on whether they give precedence to logic and evidence (Raynaud 2018) or whether they concentrate on the beliefs, social conventions, and balances of power that affect the content of the arguments and arbitration between rival rationales (Akrich et al. 2013). In any case, controversies are seen by science studies as the reflection of a social and historical reality. Disputing processes reveals balances of power, institutional positions, and social networks. They drive this social world forward by altering the balances of power, redistributing prestige and resources, and producing new conventions that will constrain future actions and positions (Lemieux 2007).

Coming back to our question—why the controversy has not taken place—the conceptual interpretations developed by Callon provide some insight (Callon 2006b, 2006a). First, the line between what warrants and what does not warrant a controversy is always the subject of agreements negotiated in the disputing

processes. In our case, the professional community of development economists gives precedence to what it considers to be fundamental research, in particular the statistical purity of randomized trials and command of causal identification biases. This aspect takes the priority here over considerations, seen as secondary by this professional community, that have more to do with an applied dimension, implying acknowledgement of the different "tricks of the trade," tactics and tweaking required to put the method into practice, and also acknowledgement of the agency of the experimenters and the trial's subjects and the sets of players it produces (Bédécarrats et al. 2019b; Kabeer 2019). This brings us back to the epistemological differences discussed above. The advent of a controversy then implies setting up sufficiently structured forums for sustained discussions to take place. In the absence of such arenas, confusion reigns among the protagonists as to who is speaking and in what context: the same players can uphold one-sided narratives in some forums and, without ever withdrawing them, make much more balanced and cautious statements in expert arenas.

0.3.1 Avoiding the Controversy, but Listening and Adapting

The absence of public dialogue does not prevent the *randomistas* from adapting their methods and practices (Ogden, Chapter 4), even though responses vary by groups of researchers. Some make their data available, thereby encouraging replications. Some acknowledge the legitimacy of methodological pluralism and combine RCTs with other methods. Some focus in detail on the impact mechanisms and processes and use specific theories (based mainly on behavioural economics). Others take the question of external validity seriously and ramp up the number of case studies in different settings (the special issue on microcredit edited by Banerjee, Karlan, and Zinman (2015) is a typical example of this; Bédécarrats, Guérin, and Roubaud, Chapter 7), or reanalyze ex-post a number of RCTs (Meager 2019). Still others take the question of "thinking small" seriously and focus on large-scale programmes and national policies. On the question of little bearing on public policies (Pritchett, Chapter 2), some *randomistas* create dedicated bodies if not become decision-makers themselves.

What remains to be seen is the extent to which the *implementation* of this new generation of RCTs in development economics can withstand contingencies on the ground and really evaluate more complex interventions. At the risk of repeating ourselves, we must emphasize the fact that one of the cruxes of the debate is this obsession with the protocol, seen as the priority over its feasibility and its ethical issues. Yet the more complicated the programmes and policies studied, the more likely it is to find tweaking, compromises made, and also risks of compromise with the initial protocol. The point is not just to adjust the technique, but to relinquish a scientistic epistemological position in the sense defined above.

0.3.2 Can We Really Afford Not to Have a Controversy Considering the Crowding-out Effects?

If a controversy is vital, it is also because the claimed hierarchy of methods has crowding-out effects, in terms of both method (the other methods are discredited), funding and types of interventions, with consequently a performative dimension: the success of RCTs is transforming the development field.

On the question of funding, consider two examples by way of illustration. In the Indian setting, a study truly capable of evaluating the impact of sanitation on infant mortality (the most appropriate indicator, but one that RCTs do not have the statistical power to capture) would cost around $90 million (subject to certain conditions; Spears, Ban, and Cumming, Chapter 6). The cost of a classic RCT is between $500,000 and $1,500,000,[21] and each RCT often generates just one published research paper. Is this cost effective when a poor country's statistical household survey system could be funded for the same amount, with a host of possible studies drawn from these observational data? This is one of the crucial questions asked by Ila Patnaik (Interviews, this volume).

On the performative effects of RCTS, the case of health is particularly illustrative. Although they may not have been the primary cause, RCTs did contribute to the rise in vertical health approaches (projects in silos) focused on the individual treatment of specific diseases at the expense of horizontal approaches designed to develop complex, integrated health systems (Garchitorena et al., Chapter 5). Other studies point up the performative (and problematic) effects of the growing use of RCTs (Adams 2016; Biehl et al. 2014): neglecting non-randomizable programmes, altering programmes to make them more easily randomizable, prioritizing evaluation at the expense of the intervention itself (in particular by changing the field staff's work (Adams 2016)). The disruption caused by RCTs and affecting the quality of interventions has been documented in other areas such as microcredit (Bédécarrats, Guérin, and Roubaud, Chapter 7) and micro-insurance (Quentin and Guérin 2013).

0.4 What Are the Research Alternatives?

Our purpose is not to reject RCTs, since they constitute a promising method … among others. However, they should still be conducted by the book, take their feasibility and ethical implications seriously by aligning with best practices established in the medical world, and interface with other methods. Although RCTs remain fit and proper for certain precisely defined policies, other methods can

[21] No precise estimate of the cost of RCTs exists to our knowledge, but Pamiès-Sumner (2015) provides approximations. See also Ravallion, Chapter 1.

and should be used, as shown by a number of chapters in this book, and the methods combined for the projects that RCTs can address (in part).

An alternative position to the gold standard is to take a pragmatic approach, defining the research questions and methodological tools required on a case-by-case basis according to the prior knowledge available, the intervention design, and the particularities of the settings, in liaison with the different stakeholders, whether field operators, donors, governments, or the largely overlooked local populations.

These alternative methods also draw on a range of methodologies based on interdisciplinarity and acknowledging the different ways of producing evidence, both quantitative and qualitative. These approaches do not set out to lay down universal laws, but to explain causal links specific to a particular time and place. Note here the disconnect between the repeated advocacy for mixed methods, whether from researchers[22] or institutions (see Rioux, Interviews, this volume),[23] and their low level of application in practice. On the side of the *randomistas*, although some publicly acknowledge the legitimacy of alternative methods (Ogden, Chapter 4), the fact that they frequently ignore the results of non-randomized methods appears to contradict this apparent open-mindedness (Bédécarrats, Guérin, and Roubaud, Chapter 7).

In the field of global health, the complexity of the interventions is such that randomization is often impracticable and observational and quasi-experimental methods are more appropriate. As shown by Garchitorena et al. (Chapter 5), there are numerous examples of alternative and complementary methods to RCTs, even if RCTs remain useful for certain specific interventions. These alternative methods have the particularity of being based on complexity theory (a health system as a whole, rather than fragmented components), combining methods and scales of analysis, drawing where possible on national statistics systems, and addressing not only impact, but also effectiveness (by introducing outputs and outcomes, but also inputs and processes into the analysis).

In addition to the examples mentioned in the book, we would also point out the need to conduct meta-analyses and replications, which are starting to emerge in development economics, but are still too thin on the ground (Camfield and Duvendack 2014). These replications can also be qualitative and revisit a field study, as has been done in Morocco and Bangladesh (Kabeer 2019; Morvant-Roux et al. 2014). Qualitative methods (semi-structured interviews, focus groups, participant observation, ethnography, case studies, life stories, etc.) can serve a number of purposes: to contextualize interventions, develop original hypotheses,

[22] See, for example, the two books mentioned in the Editors' Introduction (Cohen and Easterly 2010a; Teele 2014), wherein most of the chapters and introductory statements insist on the need for mixed methods. See also Camfield and Duvendack (2014).

[23] See Picciotto, Chapter 9, for the evaluation world in development. See also (Pamiès-Sumner 2015) for AFD, and CEDIL's work (White and Masset 2018) for DFID.

identify new and unexpected phenomena, and analyze interventions as a whole, studying the complexity of the causal links and the many, dynamic and contradictory interactions between different entities in a location-specific way. When faced with complex causal chains, which is the case with many interventions, qualitative methods are often the only way to really address the thorny question of causality (White and Masset 2018). Often (unduly) criticized for their inability to 'prove' findings, qualitative methods are also the victims of superficial and non-rigorous uses. Agnès Labrousse (Chapter 8) illustrates this misuse by discussing storytelling, a type of narrative designed to illustrate an argument, but which has no power of demonstration, which some *randomistas* misuse in their interpretations of quantitative results. At the end of the day, the only standard that holds is "good use of good evidence" (Spears, Ban, and Cumming, Chapter 6).

To sum up and wrap up this book, we believe that some key principles should guide development research, not as alternatives to RCTs, but with RCTs playing a commensurate part. These principles are probably not revolutionary…perhaps one small step for an experienced researcher, but one giant leap for humanities. First, and to make the transition from the general to the specific, research should be guided by important questions to be addressed rather than by methods for which applications need to be found. To paraphrase a famous quote: *Ask not what you can do for an RCT, ask what an RCT can do for your research*! Second, we need to get over the obsession with causal impact,[24] which has dominated the community of development economists ever since the so-called credibility revolution (Angrist and Pischke 2010). Other research questions and approaches are at least as important to advance knowledge such as analyses of observational data, thick description, analytical narratives, especially if we consider that poverty is not only a problem of deprivation but also and sometimes above all the result of social and power relations. Third, on the subject of quantitative approaches, it is essential to rebalance research efforts to take in other components of the analytic chain: what might have been gained in terms of causal attribution (in theory, since different chapters of this book show that nothing is guaranteed in practice in this area), and overinvestment in this area, has left other equally important aspects by the wayside. First and foremost, there is the question of data quality, all too often sacrificed out of a lack of interest and competence, a concern some of the most prominent *randomistas* begin to acknowledge (Dillon et al. 2020). There is then the rise in the number of replications that tackle head on a meticulous diagnostic on the data, and its inclusion in academic journals' peer review criteria. At the same time, closer attention should be paid to the question of sample designs. All too often, the implications of the use of complex sample designs are overlooked. These oversights result in the underestimation of estimator variance and the

[24] Ruhm (2019) qualifies this obsession as "the Identification police," he suggests to "shackle."

consideration of impacts as statistically significant when they are not (Gibson 2019) and when others are, but the repeatedly underpowered RCTs cannot identify them. Fourth, it is time to really put into practice two recommendations on which everyone agrees, but which remain empty talk for now without any tangible effects in practice: real consideration of the ethical issues and the combination of qualitative and quantitative methods.[25] Advocacy for Mixed Methods Approaches (MMA) is quite the combat sport.[26] Last but not least, it is time to recognize once and for all that randomized control trials are not the gold standard for evaluation. The hubris that has gripped part of the pro-RCT movement is steering research up against a brick wall (into an impasse). Restoring a sense of moderation to this immoderation is an imperative that can only do good. It is also crucial to learn from the past and acknowledge previous research, at least in two directions: the weaknesses of RCTs (Heckman, Chapter 12) and the results of non-RCT methods. If not, all aforesaid attempts to listen to the critical voices and adapt will amount, paraphrasing Lampedusa (1960) famous expression in his novel, *The Leopard*, "*everything needs to change, so everything can stay the same.*"

Will the consecration of the Sveriges Riksbank Prize lead the *randomistas* to be more balanced in their appreciation of the benefits of the different methods or, on the contrary, to take advantage of this consecration to consolidate their already virtually hegemonic position? Only time will tell,[27] but let us insist on the fact that putting an end to the "gold standard" and the quest for the "indisputable" that is characteristic of the *randomistas*' claim to superiority calls for an epistemological break, but also the advent of this controversy which we call for in earnest. Drawing on an examination of the controversies surrounding climate change, Bruno Latour (2012) advocates building debating spaces and methods to discuss and debate the different forms of scientific knowledge (in all their plurality), and non-scientific knowledge, ensuring that the ideological and political bases of these multiple forms of knowledge are neither repudiated nor eclipsed, but are spelled out and debated (Egil 2015). We believe that this project, as ambitious as it may be, is a scientific and democratic necessity if we really hope to improve development policies.

[25] As hightlighted by van der Meulen Rodgers et al. (2020) in their editorial to the *World Development* special issue on RCTs, the call for triangulation, pluralism, and collaboration (both within the scientific community and between academia, donors and civil society) is the most shared and advocated contributors' demand.

[26] This is a nod to a documentary about the work of sociologist Pierre Bourdieu, entitled *Sociology Is Quite a Combat Sport*. The documentary was directed by Pierre Carles in 2001.

[27] Based on our own experience, one may be pessimistic in this respect, given the increased difficulties in publishing critical papers in mainstream academic journals or in finding interlocutors to discuss RCTs effective contribution in the policy making arena. A mix of individual self-censorship and tided hands for institutional reasons inhibits those prone to raise their critical or nuanced voice in front of the powerful and celebrated new doxa.

0.5 Outline of the Book

The book is structured as follows. The first set of chapters presents an overview of RCTs in development (what kinds of questions can they or can they not answer?) and a range of positions on the potential of this method (Ravallion, Chapter 1; Pritchett, Chapter 2; Morduch, Chapter 3 and Ogden, Chapter 4). The second set of chapters focuses on sector analyses in health (Garchitorena et al., Chapter 5), sanitation (Spears, Ban, and Cumming, Chapter 6) and microcredit (Bédécarrats, Guérin, and Roubaud, Chapter 7), asking the following questions: what have we learnt from the RCTs in each sector and what contribution do the other methods make? The third set of chapters offers points to consider regarding the political economy, both specifically with respect to the *randomistas*' rhetoric (Labrousse, Chapter 8) and more generally by placing RCTs in the context of the field and history of development policy evaluation (Picciotto, Chapter 9). The fourth and last set of chapters expands on the proposals for improvement discussed in the sector-based chapters with a focus on specific aspects, starting with ethics, whose importance and urgency have been stressed (Abramowicz and Szafarz, Chapter 10), and then exploring statistical improvements, on the use of priors (Vivalt, Chapter 11) and non-compliance as a source of information (Heckman, Chapter 12). In the guise of an epilogue, J. Heckman offers a rereading of his critical 1992 paper, in the light of the new wave of RCTs in development. He shows that most of his conclusions, which focused on the first generation of random experiments in the field of social policy in the United States (the First awakening in his own terms), still hold. He calls to reason and to learn from the past the new generation of economists. Finally, the work ends with a section of three interviews with high-ranking policy-makers, which leaves the field of research to adopt a public policy perspective. The interviews question the use, usefulness, and responses provided by the RCTs for the decision in the real world: a first cross-interview with the CEOs of our respective institutions, specialized in the field of development: aid for the AFD (Rioux) and research for IRD (Moatti), which offers a reading seen from a northern country (France); and two interviews with high-level executives confronted daily with the development and monitoring of economic policies in India, the first field of application of RCTs in the global South (Natarajan, high Indian official; and Patnaik, former principal economic advisor to the Government of India). Upstream, with his "eleven variations," A. Deaton, in his Introduction, masterfully revisits his own insights in the light of this book's contributions.

Acknowledgement

This introduction, whose views are the sole responsibility of the editors, benefited greatly from discussions at a workshop attended by most of the authors in Paris on 17 March 2019, and from comments received from Agnès Labrousse and Ariane Szafarz whom we thank most sincerely.

Introduction: Randomization in the Tropics Revisited, a Theme and Eleven Variations

Angus Deaton

Development economists have been using randomized controlled trials (RCTs) for the best part of two decades,[1] and economists working on welfare policies in the US have been doing so for much longer. The years of experience have made the discussions richer and more nuanced, and both proponents and critics have learned from one another, at least to an extent. As is often the case, researchers seem reluctant to learn from earlier mistakes by others, and the lessons from the first wave of experiments, many of which were laid out by Jim Heckman and his collaborators[2] a quarter of a century ago, have frequently been ignored in the second wave. In this Introduction, I do not attempt to reconstruct the full range of questions that I have written about elsewhere (Deaton 2010a, Deaton and Cartwright 2018, Deaton 2010b), nor to summarize the long-running debate in economics. Instead, I focus on a few of the issues that are prominent in this volume of critical perspectives and that seem to me to bear revisiting.

The RCT is a useful tool, but I think that it is a mistake to put method ahead of substance. I have written papers using RCTs (Deaton 2012, Deaton and Stone 2016). Like other methods of investigation, they are often useful, and, like other methods, they have dangers and drawbacks. Methodological prejudice can only tie our hands. Context is always important, and we must adapt our methods to the problem at hand. It is not true that an RCT, when feasible, will always do better than an observational study. This should not be controversial, but my reading of the rhetoric in the literature suggests that the following statements might still make some uncomfortable, particularly the second: (a) RCTs are affected by the same problems of inference and estimation that economists have faced using

[1] The Nobel Prize to Abhijit Banerjee, Esther Duflo and Michael Kremer was announced as this Introduction was being revised. As it already has done, the Prize will raise the visibility of the debate about the pros and cons of conducting RCTs directed towards economic development. The extensive press discussion has revealed substantive concerns, especially about ethics. It also reveals widespread misperceptions, among both critics and defenders, about how RCTs actually work, particularly highlighting the widespread but false beliefs that randomization guarantees that the treatment and control groups are similar prior to treatment, and that an RCT can demonstrate causality.

[2] Heckman (Chapter 12, this volume), which is an updated version of Heckman (1992), and Heckman and Jeffrey A Smith (1995). See also Manski and Garfinkel (1992), which contains the 1992 version of Heckman's paper, an excellent overview introduction by Manski and Garfinkel, and several other papers that have continuing relevance.

Angus Deaton, *Introduction: Randomization in the Tropics Revisited, a Theme and Eleven Variations* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0002

other methods, as well as by some that are peculiarly their own, and (b) no RCT can ever legitimately claim to have established causality.

My *theme* is that RCTs have no special status, they have no exemption from the problems of inference that econometricians have always wrestled with, and there is nothing that they, and only they, can accomplish. Just as none of the strengths of RCTs are possessed by RCTs alone, none of their weaknesses are theirs alone, and I shall take pains to emphasize those facts. There is no gold standard. There are good studies and bad studies, and that is all. The most important things I have to say are about the ethical dangers of running RCTs in poor countries. I save those remarks for last.

I.1 Are RCTs the Best Way of Learning, or of Accumulating Useful Knowledge?

Sometimes. Sometimes not. It makes no sense to insist that any one method is best, provided only that it is feasible. It has always seemed to me to be a mistake for J-PAL to do only RCTs, and thus leave itself open to the charge that it is more (or as) interested in proselytizing for RCTs than it is in reducing poverty. Though as Tim Ogden (Chapter 4) notes, the *members* of J-PAL use a wide range of techniques in their own work, so perhaps J-PAL is just the RCT wing of a broader enterprise. Martin Ravallion (Chapter 1) is exactly right when he argues that the best method is *always* the one that yields the most convincing and relevant answers in the context at hand. We all have our preferred methods that we think are underused. My own personal favorites are cross-tabulations and graphs that stay close to the data; the hard work lies in deciding what to put into them and how to process the data to learn something that we did not know before, or that changes minds. An appropriately constructed picture or cross-tabulation can undermine the credibility of a widely believed causal story, or enhance the credibility of a new one; such evidence is more informative about causes than a paper with the word "causal" in its title. The art is in knowing what to show. But I don't insist that others should work this way too.

The imposition of a hierarchy of evidence is both dangerous and unscientific. *Dangerous* because it automatically discards evidence that may need to be considered, evidence that might be critical. Evidence from an RCT gets counted even if the population it covers is very different from the population where it is to be used, if it has only a handful of observations, if many subjects dropped out or refused to accept their assignments, or if there is no blinding and knowing you are in the experiment can be expected to change the outcome. Discounting trials for these flaws makes sense, but doesn't help if it excludes more informative non-randomized evidence. By the hierarchy, evidence without randomization is no evidence at all, or at least is not "rigorous" evidence. An observational study is

discarded even if it is well-designed, has no clear source of bias, and uses a very large sample of relevant people.

Hierarchies are *unscientific* because the profession is collectively absolved from reconciling results across studies; the observational study is taken to be wrong simply because there was no randomization. Such mindless neglect of useful knowledge is relatively rare in economics, though the failure to cite non-RCT work is common, as is its dismissal as "anecdotal" or because it is unable to separate correlation from causation (Bédécarrats, Guérin, and Roubaud, Chapter 7), but there are many worse examples in other fields, such as medicine or education. Yet economists frequently do give special weight to evidence from RCTs based on methodology alone; such studies are taken to be "credible" without reference to the details of the study or consideration of alternatives.

Economics is an open subject in the sense that good studies that produce new, important, and convincing evidence are usually judged on their merits. But it is good to be careful that merit not be a cover for methodological prejudice. When I hear arguments that RCTs have proved their worth by producing good studies, I want to be reassured that the use of randomization is not itself a measure of worth and that the argument is not circular.

I.2 Statistical Inference Is Simpler in RCTs than with Other Methods

This misunderstanding has been responsible for much mischief. One issue not often noted is that RCTs, more so than with observational research, often involve the authors in collecting data, including tracking respondents over time and recognizing and dealing with gross outliers, tasks that are far from straightforward, that involve immense amounts of time and specialized skills that not all economists possess. Problems in data gathering and handling likely dwarf the errors from mistakes in statistical inference (Bédécarrats, Guérin, and Roubaud, Chapter 7). There is nothing simple about such matters.

On inference, there are two parts to the simplicity argument. First, randomization guarantees that the two groups, treatments, and controls, are on average identical before treatment, so that any difference between them after treatment must be caused by the treatment. Second, statistical inference requires computing a p-value for the difference between two means, a simple procedure that is taught in elementary statistics classes.

Both parts of the argument are wrong.

R. A. Fisher understood from the beginning that randomization does *not* balance observations between treatments and controls, as anyone who actually runs an RCT will quickly discover. Ravallion (Chapter 1), who has long observed RCTs in the World Bank and elsewhere argues that the misunderstanding "is now

embedded in much of the public narrative" in development. It is also common in the press and in everyday parlance.

Imagine four units (villages, say), two of which are to be treated, and two not. One possibility is to let the village elders decide, for example by bidding (or bribing) to be included (or excluded), and then selecting for treatment the two villages who most want (least do not want) to be treated. This self-selection allocation of treatments and controls is clearly problematic. Yet many people seem to think that randomization fixes the self-selection. There are only six possible allocations, one of which is the self-selected allocation. We then have the absurdity that the *same* allocation is fine if it comes about randomly, but not if it is self-selected. With hundreds of villages, whether or not balance happens depends on how many factors have to be balanced, and nothing stops the actual allocation being the self-selected allocation that we would like to avoid. Nothing is *guaranteed* by randomization. Perhaps it is the idea that randomization is fair ex ante that confuses people into thinking that it is also fair ex post. But it is the ex post that matters.

Making the treatment and control groups look like one another is a good thing but requires information and *deliberate* allocation, both of which are scrambled by randomization. Fisher knew this and knew that there were more precise ways of estimating an average treatment effect by avoiding randomization, but understood that there was a difficulty in knowing what to think about the difference once measured; there will always be *some* difference even when the treatment has no effect for any unit. Randomization is a solution to this problem, because it provides the basis for making probabilistic statements about whether or not the difference arose by chance. Many years ago, the philosopher Patrick Suppes (1982) put it this way. He imagined himself presented with an urn with 50 black and white balls; there are either (A) 15 black and 35 white, or (B) 35 black and 15 white. He is allowed to draw 12 balls, and must bet on A or B. He wrote "I find it hard to imagine a sophisticated bettor who would not insist on such physical randomization before entering into the experiment." Randomization *does not* ensure balance, but it *does* allow the calculation of odds, at least in simple cases like this where nothing else affects the outcomes. Calculating odds is useful and important, but it is not the same as balance.

Many people are surprised when they are told that inference about a mean—and therefore inference about the difference between two means—is an unsolved problem. One issue was stated long ago by Bahadur and Savage (1956), who showed that without assumptions that limit skewness, the calculated t-value will generally not have the t-distribution. If we wrongly assume that it does, we will make mistakes, for example, by thinking that a large t-value indicates an effect of the treatment when, in fact, there is none. Skewness (a term that nowadays is often incorrectly used to mean bias) refers to the third moment, and in particular the presence of large outliers on one side of the distribution. Any experiment involving money is a likely example, and one can think of educational or

microfinance experiments where one or two people are immensely talented, and the others not so much (Banerjee et al. 2019).

The RAND health experiment—one of the most famous RCTs in economics—had one participant who had an immensely expensive pregnancy. In such cases, the outcome of the RCT depends on whether the outlier(s) is among the treatments or among the controls, and with an extreme enough outlier, on little else. You may think you have hundreds or thousands of observations, but in fact you only have one. Wild answers look significant, because the use of the *t*-distribution is invalidated by the skew. Trimming of outliers, or transforming the outcome variable—e.g. by taking logs—will not always help. The million-dollar baby is what will break an actual insurance scheme, however much the insurers might wish to "trim" it. We need to measure profits in dollars, not in the logarithms of dollars, let alone trimmed dollars. Perhaps the median treatment effect might be more reliable but, once again, it is the mean that breaks the budget, not the median, and even in cases where we would like to know the *median* treatment effect, it is not identified from an RCT. If you are genuinely interested in the median, you will have to use a method other than an RCT, one that requires more assumptions.

The point is *not* that RCTs have unique difficulties here, the point is that they have no exemption from such troubles, no "get out of jail free" card. Ulrich Mueller has recently shown that the problem is widespread in contemporary applied economics, particularly when using clustered robust standard errors (Mueller, 2020). When clusters are of different sizes—as in much spatial work in applied econometrics—the *p*-values that come from STATA, for example, are not reliable. My guess is that Mueller's work, which also provides a better method, will lead to substantial revisions in how we work, and in what we think we know.

In work on a related disease of inference, Alwyn Young has demonstrated that many published papers using RCTs get their *p*-values wrong (Young, 2019), so that many apparently significant results—sometimes quite startling results—are consistent with the operation of chance in a situation where the treatment has no effect. Young proposes that we return to Fisherian randomization as a way of calculating significance. If the treatment has no effect for anyone, and there is no post-randomization confounding, the estimated average treatment effect is a result only of the random allocation of subjects to treatments or controls. (Post-randomization confounding is anything other than the treatment that effects outcomes, such as "tells" in the treatment environment, or non-blinding of subjects, assessors, or analysts.) By looking at all possible random assignments in the actual data, we can tabulate the distribution of the differences in the two means under the hypothesis of no treatment effect for any unit, and calculate the probability of getting something as or more extreme than the actual difference. This "randomization inference" tests the hypothesis that the treatment has no effect for *any* individual. This hypothesis is often of interest, but it is not relevant to what we often want to know for policy, which is whether the *average* treatment effect is zero. While a zero effect

for each observation means that the average must also be zero, the converse is not true, most notably so when the treatment affects different individuals in opposite directions. A small daily dose of aspirin is an example; it saves some and kills others. In public policy, say in a teaching experiment, we might well want to know whether the new method increases test scores on average, not just whether it works for someone. (An additional complexity is that a statistical test can sometimes accept the hypothesis that each of a group of estimates is zero, but reject the hypothesis that their average is zero. Beyond that, randomization inference can itself be misled by an unfortunate sample.)

Because the calculated significance levels are unreliable in realistic situations, it is wise to be skeptical of many of the published conclusions from RCTs. *Poor Economics* (Banerjee and Duflo 2011) presents the findings of dozens of studies, many of which are interesting and important. But results that ought to be presented as estimates tend to be presented as if they are established facts. Indeed, the rhetoric of RCTs is that trials can establish the truth. They cannot. The surprising results that come out of RCTs are sometimes not results at all, and large *t*-values ought not to persuade us that they are.

I.3 RCTs Are Rigorous and Scientific

This rhetoric is rarely if ever justified. The adjectives are used as codewords for RCTs. Frequently so. The rhetoric appears to be successful, at least with funders. It is often coupled with an appeal to the importance of RCTs in medicine, but rarely coupled with a realistic reading of the successes and failings of RCTs in medicine. In the US, drugs require successful RCTs in order to be licensed, yet prescription opioids, such as OxyContin, have killed hundreds of thousands of Americans in the last twenty years. There are differences between how RCTs work in social science and in medicine, a topic on which more thinking could usefully be done. On one occasion, I discussed a series of development trials with a senior funding manager of a large foundation. He was happy to admit that the results were limited in applicability, and that some of the results were likely incorrect, but was unimpressed. RCTs, after all, he told me, are more rigorous than any other method and for him, that was enough. I think he had a notion that rigor meant that the results were generalizable, or could be scaled up. Or perhaps he held the common belief that all other methods are worse. Being wrong did not appear to conflict with being rigorous.

I.4 External Validity

"Finding out what works" is another common rhetorical slogan that, at least judged by its repetition, is effective among the public. Nothing works except in

context, and finding out what works where and under what circumstances is a real scientific endeavor. What works also depends on for whom and for what purpose; what works involves values as well as facts. There is no experiment or series of experiments that can answer such questions unconditionally. That RCTs will identify what works to eliminate global poverty is a commendable but unfounded aspiration.

A result that is true in one place, at one time, and under one set of circumstances, will typically not be true in another place, another time, or under different circumstances. What works for you may "work" for me too, except that I don't like it. Once again, these things are true of all empirical findings, no matter what method is used. No one thinks that an estimate of the average income in America will be accurate a decade from now, yet an estimate of an average treatment effect, which is also a sampling-based estimate of a mean, is often treated as if it is likely to hold elsewhere, at least in the absence of evidence to the contrary.

The practice is perhaps not very different from a long-standing practice in economics to treat elasticities as constants, as in "the" elasticity of labor supply of prime age men, or "the" price elasticity of bread. My suspicion is that those elasticities are supported by strong intuitions about the nature of the goods concerned, that most men had little choice but to work, while, once upon a time, their wives had more, that staple foods are not easily substituted for, and that the demand for small luxuries is sensitive to their prices, intuitions that were supported by many studies in many places. But this is not where we are with development today. To take Lant Pritchett's example (Chapter 2), I see no reason to suppose that if chickens are better than money in Sierra Leone, they will be better than money in Laos or, for that matter in Trenton, New Jersey, nor why, if they were better in 60 trials in 60 different places, they would be better in the 61st. And beware Bertrand Russell's chicken, who learned from hundreds of replications that when she heard the farmer's footsteps, she was about to be fed, until, on Christmas Eve, he wrings her neck. As Russell noted, the chicken could have benefited from a deeper understanding of the world around her.

Deeper understanding matters. The Gates Foundation, the largest aid donor in many areas, sees scaling up as one of its central missions, and so has seized on one or two positive results in its African agriculture initiative as evidence that "it works," and extended "it" to other farms or other countries, without any theory of why it might or might not work elsewhere (Schurman 2018). We have to face the truth that what works might be different from one farm to the next, something African farmers are likely to know, even if the experimenters do not.

It is a mistake to think of internal and external validity as twin properties that are ideally possessed by high quality studies. An RCT can be perfectly conducted using a large sample and hit the ATE on the nose. Whether it is externally valid is *not* a property of the *study* but a property of the *circumstances* in which it is to be used. There is nothing invalid about a study whose result does not apply

elsewhere. External validity is about how a study is *used*; the same study may be valid in some contexts and not in others.

There is always a temptation to take an impressive study and push it beyond its original context. This too is true of observational and experimental studies alike. Raj Chetty and his coauthors have pioneered the use of merged administrative data to describe in extraordinary detail facts about the dynamics of inequality in the United States, and have so generated huge advances in knowledge. One important finding (Chetty et al. 2019) is that, between 1989 and 2015, African American children were less likely than white children to move up the income distribution from their parents' position. Yet in many popular accounts in the press, "were" is replaced by "are," even though marriage and incarceration patterns have been changing in both groups. These are outstanding studies, among the very best in economics today, but they can make no more claim to external validity than can outstanding RCTs. Once again, the issue of external validity is general, and RCTs have no "get out of jail free" card. It may be that, without internal validity, a trial result is unlikely to hold elsewhere, but it is certainly not true that internal validity implies external validity. I do not know of explicit claims to the contrary, but I have often been struck by the contrast between the care that goes into running an RCT and the carelessness that goes into advocating the use of its results. The phrase "primacy of internal validity" can seem to justify such practices.

That the results of an RCT will be used in a context different from that in which it was done can inform the design of the trial to make it more useful. If we think that treatment effects are different in different subpopulations, then stratification by those subpopulations will not only improve the precision of the trial, but will also allow reweighting to a new situation. Scaling up will often affect potential variables that are constant across arms of the trial; for example, if an educational policy trains more students, wages are likely to fall, so that including a low wage arm of the trial might give useful information. The RCT can help provide the tools for modeling the policy consequences instead of simply leaping over or ignoring the gulf that lies between a trial and its implementation. But an RCT is unlikely to be enough by itself.

The fact that a given study replicates in different contexts in different countries—as in the study of graduation programs (Banerjee et al. 2015a) in *Science*—is indeed surprising, though it is unclear that the gains could be replicated by government workers facing realistic financial and political incentives, incentives that are quite different from those faced by highly educated graduate assistants from abroad who want the project to succeed. Yet, in such a cross-country study it is not at all clear what replication means, what measure we want to be replicated, or what we can learn from replication. We might want something like the rate of return on investment, or perhaps the fraction of people lifted above some local or global poverty threshold per unit of international currency. Instead, the authors use the "effect

size," which is the ATE standardized by the standard deviation of the treatment. In the words of Arthur Goldberger and Charles Manski (1995: 769), "standardization accomplishes nothing except to give the quantities in noncomparable units the superficial appearance of being in comparable units. This accomplishment is worse than useless—it yields misleading inferences."

I.5 Pre-registration of Trials

I unsuccessfully argued against the American Economic Association (AEA) requiring pre-registration of the trials whose results are to be published in its journals. I think it is a bad idea for the AEA to legislate on methods rather than assessing studies on their merits. In my experience as an economist and while serving on AEA committees, disagreements between economists that are, in truth, political or personal, are often presented as methodological differences. The AEA has, at least since the 1930s, been successful in avoiding schisms and has remained a broad church for economists of all stripes, and its presidents have ranged from Milton Friedman to Kenneth Galbraith, though I doubt they thought much of each other's methods. (Friedman tried unsuccessfully to block Galbraith's presidency.)

The problems of *p*-hacking, data mining, and specification searches are real enough. Funders who have spent large sums on an RCT often exert pressure to find at least one subgroup for which the treatment was effective. But, once again, such problems are not specific to RCTs. Some have indeed argued for preregistration for *all* studies, so that, before I start work on an observational study using the census, for example, I should notify the AEA—or perhaps the Census Bureau—of my data analysis plan. It is not clear where all this stops; must I report a conversation with a colleague or a finding that I read about in the newspaper that shapes my agenda or limits my choice of variables?

The findings of my own of which I am most proud have all had a large element of serendipity, though I was informed enough to know what I was looking *at*, even when I was looking *for* something else. None of these results would have appeared in a pre-analysis plan and would thus not be publishable in the *Journal of Correctly Done Studies*. Bill Easterly has noted that Columbus could not have discovered America if he had been required to stick to a pre-analysis plan filed in a lockbox in Seville or Genoa (Easterly 2012). I find it hard to believe that what Anne Case and I found on midlife mortality rates (Case and Deaton 2015), results that were totally unexpected to us, came from data snooping. Though I can easily imagine a statistically blinkered editor rejecting the paper because we could not produce the certificate of preregistration that authorized our work on midlife mortality. The risk of stifling important but unexpected results is surely much worse than the risk of promoting fallacious ones.

I.6 Experimentation: Kick It and See

I am all for experimentation (Morduch, Chapter 3). But there is no logical connection between experimenting and randomizing. Indeed, one might be wise, when directing one's kick, to be rather precise about one's aim; kicking at random is not advisable, and it might hurt. Randomization is about judging the significance of what has happened, not about designing a kick. The serious point here is that, in many cases, randomization is unhelpful for experimentation, it can turn a good experiment into a useless one. Information that we should be using to improve our study is scrambled.

The key laboratory experiments in economics did not use randomization (Svorenčik 2015). The Industrial Revolution is often described as having come about by endless tinkering, not by randomization, which would have got in the way of purposeful trial and error. Another example I have used in the past (Deaton 2012) is the arcade video game, *Angry Birds*. The birds need to be fired at an angle from a catapult, and can sometimes be redirected, speeded up, or detonated in flight, the object being to kill the egg-stealing pigs that are hiding in inaccessible places. Given the immense number of combinations, a systematic set of RCTs would take unimaginably long, although a dexterous child can figure out the solution in minutes. There are many kinds of experiments where randomization is not required, or would obscure the results. Randomization, after all, is *random* and searching for solutions at random is inefficient because it considers so many irrelevant possibilities, just as it did in Fisher's fields.

I.7 RCTs and Other Methods

In many discussions of RCTs, comparisons are drawn with other methods, typically instrumental variables (IV), regression discontinuity (RD) or difference in difference methods. But this is much too narrow a comparison. As someone who has lived with, used, and taught econometric methods for more than forty years, I watched the progression that led to RCTs. We used to run regressions of y on x, with much too little discussion of what generated the variation in x. We learned about differences in differences, instrumental variables and regression discontinuity as methods for purging unwanted variance from x, and creating two groups that were deemed to be identical apart from treatment. RCTs could be thought of as cleaner versions of IV, RD, or differences in differences, effectively reverting to regression but with a guaranteed assumption that x was randomly assigned. Given this history, we can see why an RCT seemed like the ultimate solution, as indeed it is when we think this way.

But as John Stuart Mill noted long ago (1843), the "method of differences," which compares two groups, one treated, one not, is only one among many ways of making causal inference. Finding out the cause of a plane crash does not involve differences (or at least we might hope not), and the hypothetico-deductive method, which is how physicists say they work, does not involve differences, simply the making and checking of predictions. That is why graphs and cross-tabulations can be so powerful when they arrange data in a way that contradicts a mass of prior understanding about how the world works. More formally, the Cowles Commission developed a method of building causal models with careful attention to mechanisms, and with a language that emphasized causal structure and procedures for delineating which parts of the structure could or could not be estimated from data. These models could be interrogated to test their predictions and the adequacy of the causal structure. Economists once used these methods more than they do today, and they comprised the main content of econometrics texts for many years, but my guess is that most graduate students in economics today would be hard pressed to define structural and reduced forms. Papers had a theory section, which developed checkable predictions, ideally predictions that are surprising and unique to the theory, which are checked out in the empirical section. Some of these methods can be interpreted as looking at differences between groups, but not all.

I.8 Small versus Large

Lant Pritchett has provided a typically eloquent, funny, and passionate argument that it is growth that matters for poverty reduction, not "rigorous" (or not) project by project evaluation, whether of money or chickens (Pritchett, Chapter 2). In *Poor Economics*, Abhijit Banerjee and Esther Duflo argue the opposite, that it is only at the level of the "small" that we know what we are doing, so we must build knowledge trial by randomized trial.

The debate is (at least) as old as the World Bank. Here is a simplified history. The Bank started out with the small, doing projects, ports, roads, power plants, and the like. It became quickly obvious that evaluating projects using commercial criteria often did not improve people's lives, particularly in economies where prices were distorted by tariffs, marketing boards, rationing, or exchange controls. An early response by two groups of very distinguished economists was to develop shadow prices to replace the market prices. Partha Dasgupta, Stephen Marglin, and Amartya Sen (1972) produced one set of methods for the United Nations, and Ian Little and James Mirrlees (1974) another for the OECD. The latter was turned into a manual by Lyn Squire and Herman van der Tak (1975) for use in the World Bank. Yet the calculations were sometimes elaborate, beyond the capabilities or

inclinations of bank lending officials whose own incentives were to move money quickly. And the rules must have seemed incomprehensible to policy-makers in the countries asked to implement them. As an example of the primitive state of project evaluation in much of the world, Lyn Squire later noted (1989: 1126–7) that even the most elementary tool of project evaluation, the discounting of future benefits, was rarely used in borrowing countries left to themselves. (This was not the case in India, where economists in the Planning Commission meticulously calculated shadow prices with at least some swallowing their personal skepticism.) If the economy was comprehensively distorted, there was surely little point in evaluating projects at market prices, and evaluation at shadow prices was not a feasible alternative.

The remedy was to switch from the small to the large, to fix the distortions first, and to get the macroeconomy right before doing project evaluation. Structural adjustment was the result.

In support of this, empirical analyses, like Pritchett's, showed that economic growth was the way to generate material poverty reduction. The great episodes of material poverty reduction in the world—particularly China and India—were driven by economic growth and by globalization. Aggregate growth came with growth in the small too, more jobs, more opportunities, more roads, more and better schools and clinics, but those were seen as springing up more or less spontaneously in an economy with good institutions and where rapid growth was ongoing. None of this explained *how* to stimulate economic growth. For this, cross-country regressions were seen as a help. These were widely criticized and are easily mocked, but yielded some useful knowledge, such as the importance of domestic investment—certainly a key in China, India, or Korea—of the provision of public goods, and that foreign aid, even at its best, was not likely to do much to stimulate growth by itself. They also systematized and disciplined the evidence, which was better than the country by country anecdotes (aka war stories) that had dominated much of the previous discussion. But we learned more about what slows growth than what speeds it up. All valuable, but hardly the keys to eliminating poverty through faster growth. No one, as far as I am aware, suggested that RCTs were the key to economic growth; it is hard to tell a story in which RCTs had any relevance for poverty reduction in China (Yang, 2019).

The Bank was half right. The better macroeconomic management in many countries around the world, the better understanding of monetary policy and central banking, as well as of the costs of exchange rate undervaluation and commodity price taxation have all contributed to better growth and poverty reduction, especially given time to operate (Easterly 2019). Economists today, adherents of the credibility revolution and of testing for causality, tend to dismiss such evidence on the grounds that, in their view, it is neither rigorous nor credible. (Yet they have no similar difficulty with the causal claim that RCTs are effective in reducing global poverty.)

Those who believe that external help can aid economic development need to square the circle. No one doubts the importance of the macro perspective, only that the tools to influence economic growth are limited. The micro level trials are often successful in themselves, but their role in diminishing poverty rates is largely a matter of faith. RCTs need a theory of implementation, or of scale up, that explains just how the results are to be used in practice. That has to include attention to unintended consequences—the effects of implementation on the actions of government and communities—that are not usually included in the end-points of the trial. General equilibrium effects need to be thought through; scaling up will change prices and behaviors that were held constant in the experiments. RCTs routinely make the assumption that spill-over effects do not exist (the SUTVA assumption), yet the assumption is routinely violated, for example in sanitation (Spears, Ban, and Cumming, Chapter 6) or deworming projects. At the individual level, the treatment works and spillovers on others are small and often cannot be (or are not) measured. Yet, at the aggregate level, the sum of the individually small spillovers can negate or reverse the effect.

I.9 Models

There is a great attraction of being able to make policy recommendations without having to construct models. I understand the appeal, of allowing the data to speak, or of generating data that speak for themselves, but I believe that attempts to do so are bound to fail. Interpreting an RCT always requires assumptions. We need to assume that it is only the treatment that matters, which is impossible to guarantee without careful policing of post-randomization confounding, just as it is impossible to be sure that the exclusion restrictions are valid for estimation using instrumental variables. People do not always accept their assignment, which can be handled by using intent to treat estimation, though the intent-to-treat average treatment effect is often not what we need to know. Or we can build models of why people do or do not accept their assignments, which is in itself potentially useful information (Heckman and Smith 1998). What happens if an RCT gives a positive effect when the outcome is measured in levels, but a zero effect when measured in logarithms? Such cases are easy to construct.[3]

As practitioners are aware, the use of prior information will improve precision (Vivalt, Chapter 11). In practice, average treatment effects are often estimated by running a regression that includes control variables. These have to be chosen, and

[3] Consider a small binary treatment that changes log income by an amount a that varies over units, but averages to zero. The effect on individual income is $a.y$, where y is income. The mean of $a.y$ depends on the correlation of income and the individual treatment effect, which can be positive, negative, or zero.

it is not clear by what rules variables are included or excluded, or how many to use. Stratification can increase precision too, but only if the stratification uses valid prior information about differences in the average treatment effects across strata.

The *use* of trial results is where modelling becomes essential. We need some theory to tell us whether the results have some relevance elsewhere, and if so, how to adapt them.

I.10 Causality

A well-designed RCT will tell us *something* about causality. Yet, once again, there are many assumptions that need to be made to get from the data to the conclusion. In any finite trial, and there are no others, the possibility that the result is due to chance can never be ruled out. The measurement of the outcomes may matter, as in the example of levels versus logarithms. To quote the philosopher and epidemiologists Alex Broadbent, Jan Vandenbroucke, and Neil Pearce (2017: 1844), "Causal conclusions do not follow deductively from data without a strong set of auxiliary assumptions, and these assumptions are themselves not deductive consequences of the data." In the same paper they write, "we suggest that it is good practice to refrain from calling any individual study's estimate 'causal' even if it is a randomized trial. It is the totality of the evidence that leads to the verdict of causality. Causality is a scientific conclusion, a *theoretical* claim, and as such transcends any individual study" (italics added). Causality is in the mind, not the data, an idea that Heckman and Pinto trace back to Frisch and Haavelmo (Heckman and Pinto 2015). The triangulation of results, or learning about causal processes from many studies over time, is well-illustrated by the chapter on sanitation in this volume (Spears, Ban, and Cumming, Chapter 6).

It is worth noting that it is not just the results of an RCT that may fail to transport, but causality itself. Nancy Cartwright and Jeremy Hardie (2012) illustrate with a Rube Goldberg machine in which opening a window leads, through a long chain of preposterous but effective causal connections, to a pencil being sharpened by a woodpecker. Yet opening windows does not usually sharpen pencils, and a causal chain in one setting may be quite different in another setting. My impression is that when economists put the word "causal" in the titles of their paper, they are claiming more than a single instance in a specific context. Beware of Rube Goldberg.

That there are other ways of building causal models is well-known to economics students brought up in the Cowles tradition, or to readers of Judea Pearl (Pearl and Mackenzie 2018). Pearl argues that we have to *start* with a causal model and then use it to confront the data and to test its structure and, like the Cowles Commission before him, offers a series of tools and methods to do so. The wisdom of Austin Bradford-Hill's discussion (1965) of the many ways to detect

causality seems to be little referred to in economics; Bradford-Hill was the pioneer in randomized clinical trials seventy years ago and it sometimes seems as if we are losing, wisdom not gaining it.

I.11 Ethics

It is good that economists should think about the ethics of experimentation. I have very little to add to the discussions about equipoise and informed consent that are covered elsewhere in this volume (Abramowicz and Szafarz, Chapter 10). Yet some of the development RCTs seem to pose challenges to the most basic rules. How is informed consent handled when people do not even know they are part of an experiment? Beneficence is one of the basic requirements of experimentation on human subjects. But beneficence for whom? Foreign experimenters or even local government officials are often poor judges of what people want. Thinking you know what is good for other people is not an appropriate basis for beneficence.

Ethics also require us to be realistic about what RCTs can and cannot do. Ethical lapses are more easily justified for those who subscribe to the hierarchy view, that the only evidence that counts is evidence from RCTs, thus ruling out options that might pose fewer risks to subjects and might lead to better conclusions. Telling developing country policy-makers that RCTs are the only way of gathering evidence for policy is unethical, because it can cause them to ignore important information. The previously discussed issues of getting *p*-values right is relevant here too. An underpowered trial that cannot possibly establish its aims is also unethical when it imposes burdens on subjects.

My main concern is broader. Even in the US, nearly all RCTs on the welfare system are RCTs done *by* better-heeled, better-educated, and paler people *on* lower income, less-educated and darker people. My reading of the literature is that a large majority of American experiments were not done in the interests of the poor people who were their subjects, but in the interests of rich people (or at least taxpayers) who had accepted, sometimes reluctantly, an obligation to prevent the worst of poverty, and wanted to minimize the cost of doing so (Gueron and Rolston 2013). That is bad enough, but at least the domestic poor get to vote, and are part of the society in which taxpayers live and welfare operates, so that there is a feedback from them to their benefactors. Not so in economic development, where those being aided have no influence over the donors. Some of the RCTs done by Western economists on extremely poor people in India, and that were vetted by American institutional review boards, appear unethical, sometimes even bordering on illegality, and likely could not have been done on American subjects (Sarin 2019). It is particularly worrying if the research addresses questions in economics that appear to have no potential benefit for the

subjects. Using poor people to build a professional CV should not be accepted. Institutional review boards in the US have special protection for prisoners, whose autonomy is compromised; there appears to be no similar protection for some of the poorest people in the world. There is an uncomfortable parallel here with the debates about pharmaceutical countries testing drugs in Africa.

I see RCTs as part of what Bill Easterly (2013) calls the "technocratic illusion," that is the original sin of economic development, an aspect of what James Scott (1998) has called "high modernism," that technical knowledge, even in the absence of full democratic participation, can solve social problems. According to this doctrine, which seems especially prevalent in Silicon Valley, among foundations, and in the effective altruism movement, global poverty will yield to the right technical fixes, one of which is the adoption of RCTs as the basis for evidence-based policy. Ignoring politics is seen as a virtue, not the vice that it is. Foundations and altruists often "know" what is good for poor people, and have the best intentions, but provide little evidence that poor people agree with their assessments or value their remedies, so that their interests can easily come to conflict with those they are trying to help. The technocrats believe that they can develop other people's countries from the outside, because they know how to find out what works. In this, at least, there is no great difference between designing a gadget and designing social policy. Both are exercises for engineers.

Engineering poverty reduction is at best hopeless, and at worst disastrous. Development agencies today use the word "partnership" a great deal, but there is no genuine partnership when all the money is on one side. Nor can there be genuine informed consent in an RCT when aid money is at stake.

Finding out what works is not the same thing as finding out what is desirable. Good intentions by donors are no guarantee of desirability. Jean Drèze (2018a) has provided an excellent discussion of the issues of going from evidence to policy. One of his examples is the provision of eggs to schoolchildren in India, a country where many children are inadequately nourished. An RCT could be used to establish that children provided with eggs come to school more often, learn more, and are better nourished. For many donors and RCT advocates, that would be enough to push for a "school eggs" policy. But policy depends on many other things; there is a powerful vegetarian lobby that will oppose it, there is a poultry industry that will lobby, and another group that will claim that their powdered eggs—or even their patented egg substitute—will do better still. Dealing with such questions is not the territory of the experimenters, but of politicians, and of the many others with expertise in policy administration. Social plumbing should be left to social plumbers, not experimental economists who have no special knowledge, and no legitimacy at all (Duflo 2017).

Working to benefit the citizens of other countries is fraught with difficulties. In countries ruled by regimes that do not care about the welfare of their

citizens—extractive regimes that see their citizens as source of plunder—the regime, if it has complete control, will necessarily be the beneficiary of aid from abroad. This is most obvious in war zones where it is impossible to deliver aid without paying off the warmongers and prolonging or worsening the suffering (de Waal 1997). The dilemma extends to peacetime too. In authoritarian regimes with full control, it is only possible for outsiders to help when it is in the government's interest to accept that help. Development agencies then find themselves in the situation of being "allowed" to help the poor, or to help provide health services, while providing political cover for the "enlightened" despot who is thereby free to persecute or eliminate his opponents (Deaton 2015). Similar issues arise in democracies too, though less sharply; the step from evidence to policy is never ethically neutral but is less fraught when the poor have a voice and some political power.

What does this have to do with RCTs? Irrelevance for one. It makes no sense to spend resources randomizing schools or medicines when the President, facing an election, is imprisoning his foes or inciting violence against his tribal and political enemies (Wrong 2009). As larger numbers of the world's poor come to live in nominally democratic states with populist autocratic leaders, more and more ethical dilemmas will confront trialists. Why are agencies funding aid, or RCTs to support aid, in countries whose leaders do not accept the liberal democratic beliefs of the donors and experimenters? I am not claiming there are no answers to this question, only that donors need to know what they are.

There have already been protests[4] about the Bill & Melinda Gates Foundation's award of one of its Global Goal Awards to Narendra Modi for building toilets in India, at a time when Modi is depriving Kashmiris of their rights, is threatening to remove citizenship from millions of Assamese, and is showing a preference to use religion as a criterion to confer citizenship on immigrants. The Foundation argues that the reward recognizes only Modi's achievements in sanitation; this is surely a perfect example of limitations and dangers of technocratic aid. It empowers despotism and intolerance. Modi has received other prestigious awards from development agencies, including the United Nations. And much worse has happened repeatedly in Africa.

Aid agencies are turning a blind eye to political repression so long as the oppressors help check off one the Sustainable Development Goals, preferably as demonstrated by randomized controlled trials. The RCT is in itself a neutral statistical tool but as Dean Spears notes[5], "RCTs provide a ready and high-status

[4] Hamid, Sabah (2019). "Why I resigned from the Gates Foundation," *New York Times*, September 26. "Dismay at Gates Foundation prize for Narendra Modi," *The Guardian*, Letter, 23 September, 2019, "Bill and Melinda Gates Foundation under fire for award to Narendra Modi," *The Guardian*, 12 September 2019.

[5] Dean Spears, personal communication, October 14, 2019. Quoted with permission.

language" that allows "mutual legitimization among funders, researchers, and governments." When the RCT methodology is used as a tool for "finding out what works," in a way that does not include freedom in its definition of what works, then it risks supporting oppression.

Acknowledgements

For (generous, helpful, and amazingly rapid) comments on an earlier version, I am most grateful to Nancy Cartwright, Anne Case, Shoumitro Chatterjee, Nicolas Côté, Jean Drèze, William Easterly, Reetika Khera, Lant Pritchett, Dean Spears, and Bastian Steuwer. I acknowledge financial support from the National Institute for Aging through NBER, Grant No. P01AG05842.

1
Should the Randomistas (Continue to) Rule?

Martin Ravallion

1.1 Introduction

The new millennium has seen a huge increase in the application of impact evaluations to developing countries, typically with the aim of improving policy-making. The International Initiative for Impact Evaluation (3ie) has compiled metadata on such evaluations, as reproduced in Figure 1.1.[1] We see a remarkable 30-fold increase in the annual production of published IEs since 2000, compared to the 19 years prior to 2000.[2]

There are two broad groups of methods, as also identified in Figure 1.1.[3] In the first, access to a program (the "treatment") is randomly assigned to some units, with others randomly set aside as controls. To measure the program's impact one then compares mean outcomes for these two samples. This is the simplest version of a randomized controlled trial (RCT). A second group of methods does not use randomization. These include purely "observational studies" in which the assignment of treatment is taken as data, and understood to be purposive rather than random. The second group also includes deterministic assignments, such as based on priors about likely benefits from the treatment. While some non-RCTs that help inform policy-making are purely descriptive, others attempt to control for the pre-treatment differences between treated and untreated units based on what can be observed in data, with the aim of drawing credible causal inferences about impact.

While the use of RCTs in development applications began around 1980, a rapid expansion in their use emerged some 20 years later. About 60 percent of the impact evaluations since 2000 have used randomization. The latest 3ie count has

[1] See Cameron et al. (2016) and Sabet and Brown (2018). The numbers here span 1981–2015.

[2] 4501 impact evaluations are recorded in the 3ie database, covering the period 1981–2015, of which 4338 were published in 2000–2015. The annual rates are 271 since 2000 and 9 for 1981–1999.

[3] The 3ie series is constructed by searching for selected keywords in digitized texts. 3ie staff warned me (in correspondence) that their old search protocols were probably less effective in picking up observational studies relative to RCTs prior to 2000. So the earlier, lower, counts of non-randomized evaluations in Figure 1.1 may be deceptive. The 3ie counts pick up many more RCTs than reported in Bouguen et al. (2019) (also noting that the latter give cumulative totals not annual flows).

Martin Ravallion, *Should the Randomistas (Continue to) Rule?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective.* Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0003

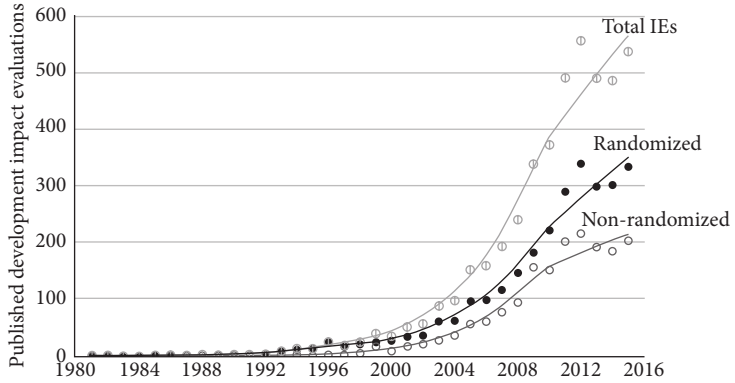

Figure 1.1 Annual counts of published impact evaluations for developing countries

Note: Fitted lines are nearest neighbor smoothed scatter plots. See footnote 3 in the main text on likely under-counting of non-randomized evaluations in earlier years. Source of primary data: International Initiative for Impact Evaluation.

Source: Author, based on 3ie data.

333 papers using this tool for 2015.[4] The growth rate is striking. Fitting an exponential trend (and the fit is good) to the counts of RCTs in Figure 1.1 yields an annual growth rate of around 20 percent—more than double the growth rate for all scientific publishing post-WW2.[5] As a further indication, if one enters "RCT" or "randomized controlled trial" in the Google Ngram Viewer one finds that the incidence of these words (as a share of all ngrams in digitized texts) has tended to rise over time and is higher at the end of the available time series (2008) than ever before.

The fact that so much of the growth evident in Figure 1.1 has been for RCTs would surely not have been anticipated prior to 2000. After all, RCTs are not feasible for many of the things that governments and others do in the name of development. Nor had RCTs been historically popular, given the often-heard concerns about withholding a program from some people who need it, while providing it to some who do not, for the purpose of research. Development RCTs used to be a hard sell. Something changed. How did RCTs become so popular? And is their popularity justified?

Advocates of RCTs have been dubbed the "randomistas."[6] They proffer RCTs as the "gold standard" for impact evaluation—the most "scientific" or "rigorous"

[4] To put this in perspective for economists, this is about the same as the total number of papers (in all fields) published per year in the *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Econometrica* and the *Review of Economic Studies* (Card and DellaVigna 2013).

[5] Regressing the log RCT count (dropping three zeros in the 1980s) on time gives a coefficient of 0.20 (s.e.=0.01; n=32; R^2=0.96) or 0.18 (0.01; n=16; R^2=0.96) if one only uses the series from 2000 onwards. In modern times (post-WW2), the growth rate of scientific publications is estimated to be 8–9% per annum (Bornmann and Mutz 2014).

[6] That term "randomistas" is not pejorative; indeed, RCT advocates also use it approvingly, such as Leigh (2018).

approach, promising to deliver largely atheoretical and assumption-free, yet reliable, IE.[7] This view has come from prominent academic economists, and it has permeated the popular discourse, with discernable influence in the media, development agencies and donors, as well as among researchers and their employers.[8] This is an unconditional preference for RCTs. While there are a great many contexts for an impact evaluation (types of interventions, sectors of the economy, countries, communities, social/ethnic groups), the gold-standard claim is typically made independently of context.

There have been pushbacks. RCTs in social-policy applications have raised many concerns.[9] Critics have argued that (inter alia): the assumptions required for a reliable impact estimate using an RCT need not hold in reality; RCTs are ethically questionable; and the "black box" nature of RCTs limits their usefulness for policy-making, including both scaling up and learning about likely impact in other contexts. There have been defenses against the critics.[10] And passions have run high at times, with one commentator dismissing ethical criticisms of RCTs as "garbage" (Fiennes 2018) while one critic has called the RCT revolution "madness" (indeed, "much worse than madness" at one point) (Pritchett 2020).

In the light of the rising prominence of development RCTs, and the continuing debates, this chapter returns, ten years later, to the question posed in Ravallion (2009a), "Should the randomistas rule?" The sense in which randomistas "rule" is in their claimed hierarchy of methods, which is the foundation of their intellectual authority and power to persuade.[11] That hierarchy is the main focus of this chapter. While recognizing the attractions of RCTs for some purposes, the chapter argues that the supportive public narrative on RCTs that has emerged is not well grounded in an appreciation of the limitations of this research tool. The chapter's intended audience is not the experts on either side, but the broader

[7] For example, Banerjee (2006) writes that: "Randomized trials like these—that is, trials in which the intervention is assigned randomly—are the simplest and best way of assessing the impact of a program." Similarly, Imbens (2010: 407) claims that "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top." And Duflo (2017: 3) refers to RCTs the "tool of choice."

[8] An example of the broader influence of the "gold standard" view is the Wikipedia entry on IE, which states that "Randomized field experiments are the strongest research designs for assessing program impact . . . as it allows for a fair and accurate estimate of the program's actual effects." In another example, Keating (2014) writes that "Randomistas, proponents of randomized controlled trials, have recently been transforming the way we think about economic development and aid to poor countries." Similarly, Leigh's (2018) volume is entitled *Randomistas: How Radical Researchers Changed Our World*.

[9] See Heckman and Smith (1995), Grossman and Mackenzie (2005), Cartwright (2007), Ravallion (2009a,b, 2012), Rodrik (2009), Barrett and Carter (2010), Deaton (2010a), Keane (2010), Baele (2013), Basu (2014), Mulligan (2014), Pritchett and Sandefur (2015), Favereau (2016), Ziliak and Teather-Posadas (2016), Hammer (2017), Deaton and Cartwright (2018), Gibson (2019), Pritchett (2020) and Young (2019).

[10] Including Banerjee and Duflo (2009), Goldberg (2014), Imbens (2010, 2018), Glennerster and Powers (2016) and McKenzie (2019).

[11] Thus, McKenzie's (2019) observation that only 10 percent of all papers in development economics (any field, in 14 journals) are RCTs does not refute the claim that the randomistas do indeed rule in the sense used here.

community of economists and other social scientists, donors, policy-makers and their advisors, students, and young researchers.

The chapter begins with an overview of the theory of impact evaluation, as relevant to the choice of methods (Section 1.2). It then discusses the randomistas' influence on development research (Section 1.3), the concerns about the ethical validity of their preferred method (Section 1.4), and the relevance of their research to policy (Section 1.5). Section 1.6 concludes.

1.2 Foundations of Impact Evaluation

The focus is on assigned programs, in that some units (the "treated") in a well-defined population get the program and some do not. Imagine drawing two random samples from the population, one from those treated and one from those not, and then measuring relevant outcomes for both. This constitutes a single experimental trial.[12] The difference in mean outcomes is the trial's estimate of the true mean impact for that population, also called the average treatment effect (ATE). That estimate can differ from the true value due to measurement errors, sampling variability, spillover effects ("contamination") between the two groups, monitoring effects, and/or systematic bias arising from any confounding variables that jointly alter outcomes and treatment status. Each trial's sampled pair gives a different estimate, sometimes too high, sometimes too low, though we never know by how much since we do not (of course) know the true value. Every trial has some error.

The ideal RCT is the special case of the above setup in which the trial's treatment status is also assigned randomly (in addition to drawing random samples from the two populations, one treated and one not) and the only error is due to sampling variability. This ideal can be illusive in practice, especially with human subjects; the discussion will return to the real-world departures from the ideal, but for now an ideal RCT is assumed. In this special case, as the number of trials increases, the mean of the trial estimates tends to get closer to the true mean impact. This is the sense in which an ideal RCT is said to be unbiased, namely that the experimental error is driven to zero *in expectation*. This property also allows us to estimate the variance of the estimates. Thus, using both random treatment and random sampling facilitates calculation of the standard error of the impact estimate from an RCT, to establish a statistical confidence interval.[13]

[12] The word "experiment" is sometimes defined as any situation in which the evaluator controls everything, and this is deemed to be the case for an RCT; see, for example, Cox and Reid (2000). However, it is almost never the case that the evaluator controls everything in RCTs with human subjects, as used to evaluate social policies. Here I use the broader definition of "experiment," not assuming full control. It may or may not be an RCT.

[13] Current practices in this respect can be questioned. Young (2019) points to a number of concerns in past impact estimates of standard errors when using RCTs with regression controls and shows that many published economics papers have over-estimated the statistical significance of their impact estimates. Also see the discussions in Deaton and Cartwright (2018) and Imbens (2018).

Prominent randomistas have sometimes left out the "in expectation" qualifier, or ignored its implications for the existence of experimental errors (as noted by Deaton and Cartwright, 2018). These advocates of RCTs attribute *any* difference in mean outcomes between the treatment and control samples to the intervention.[14] This common mistake might be thought of as little more than a minor expository simplification.[15] However, the simplification is now embedded in much of the public narrative. Beyond the experts (putting aside their unguarded statements), many people in the development community now think that any measured difference between the treatment and control groups in an RCT is attributable to the treatment. It is not; even the ideal RCT has some unknown experimental error.

A rare but instructive case is when there is no treatment. Absent any other effects of assignment (such as from monitoring), the impact is zero. Yet the random error in one trial can still yield a non-zero mean impact from an RCT. An example is an RCT in Denmark in which 860 elderly people were randomly and unknowingly divided into treatment and control groups prior to an 18-month period without any actual intervention (Vass, 2010). A statistically significant (prob. = 0.003) difference in mortality rates emerged at the end of the period.

In the light of these observations, consider the choice of methods. Suppose that, with a given budget, we can implement either an RCT or an observational study. For the latter, people select into the program, and we take random samples of those who do and those that do not. We want to rank the methods ex ante according to how close their trial estimates are likely to be to the true value. Let us say that an estimate is "close to the truth" if it is within some fixed interval centered on the true value. (The focus here is on the "internal validity" of each estimator—its accuracy for the population in hand; Section 1.5 turns to "external validity.")

The reason one hears most often for the "gold-standard" ranking is the unbiasedness of an ideal RCT. Economists have focused a lot on one particular source of bias, namely any difference between the mathematical expectation of a parameter estimate and its (unknown) true value. (In some of the literature this is called "systematic bias," as distinct from the, potentially many, sources of trial-specific

[14] For example, with reference to RCTs, Banerjee and Duflo (2017) write that "any difference between the treatment group and the comparison group can be confidently attributed to the treatment," and Banerjee et al. (2019) claim that an RCT ensures that "any difference between treatment and control units reflects the impact of the treatment." One finds a similarly unguarded claim in the "Introduction to Evaluation" on the website of the Abdul Latif Jameel Poverty Action Lab (J-PAL) (which Section 1.3 returns to); having described a stylized RCT for a water purification project, with treatment and control groups, J-PAL says that: "any differences seen later on can be attributed to one having been given the water purification program, and the other not." Another example is found in a technical manual on impact evaluation by the Inter-American Development Bank and the World Bank (Gertler et al. 2016).

[15] As Imbens (2018) suggests, in his comments on Deaton and Cartwright (2018).

errors.[16]) Even by this narrow definition, an observational study need not be biased. One typically adjusts for covariate imbalance, including in an RCT. Bias is removed when the treatment status is conditionally exogenous, i.e., uncorrelated with the error term conditional on the covariates (though this is clearly a stronger assumption than for an RCT). That assumption may or may not be acceptable, depending on the context (the program and the data available). Whether or not the treatment is exogenous given the control variables depends on whether those variables adequately reflect the determinants of treatment placement; that must be judged in each setting. A good understanding of the economic and social determinants of program placement—the decision problems facing the various stakeholders in the specific context—can help in determining what data one needs. Omitted confounders will often remain, although that need not mean large biases on adjusting for the observed confounders.

If unmeasured confounders are a serious concern then the bias can be removed if one can find a source of exogenous variation in treatment status that is not also a determinant of outcomes given treatment. This is an instrumental variable (IV). A valid IV must be correlated with treatment status *and* uncorrelated with outcomes, given treatment and the control variables. In a regression, this requires that the IV is uncorrelated with the error term—giving what is called the "exclusion restriction." This condition must ultimately be judged on theoretical grounds, though close study of the factors determining treatment status in the specific setting can be valuable in finding theoretically plausible IVs, as well as potential confounders. For example, consider a program for which the assignment to treatment depends on whether an eligibility score is above some critical threshold, along with other factors adding fuzziness to the assignment. As long as the threshold is arbitrary (namely that mean counterfactual outcomes do not change at the threshold), whether the score is above or below this critical value is a theoretically defensible IV.[17] Though less familiar to economists, bias in an observational study due to unmeasured confounders can also be eliminated if there is an intermediate variable that links treatment to outcomes but does not depend on the confounders.[18]

Even if we agree that an RCT is better at removing bias in a specific setting, that does not clinch the ranking. There are two main reasons. First, given the constraints faced on RCTs in practice, it may not be feasible to properly represent the population of interest. At least when there is a free media, governments are likely to see a political risk in supporting ethically questionable research. While RCTs

[16] On the multiple sources of "bias" see Hernán and Robins (2018).

[17] This is an example of regression-discontinuity design; for a formal treatment see Hahn, Todd, and Van der Klaauw (2001).

[18] This is sometimes called "front-door adjustment" as distinct from "back-door adjustment" using an IV (Pearl and Mackenzie 2018, Chapters 4 and 7). An example of front-door adjustment can be found in Glynn and Kashin (2018). For a more formal treatment see Pearl (2009, Chapter 3).

are sometimes done with governments, more benign observational studies are often easier to accept. Thus, academic randomistas looking for local partners see the attractions of working instead with local non-governmental organizations (NGOs). The desire to randomize may thus deliver (under ideal conditions) an unbiased impact estimate for a non-randomly selected subpopulation, such as those in the catchment area of a cooperative local NGO. Furthermore, the selection process for the compliant subsample may be far from clear (indeed, without even a mention in the paper on how it was chosen). It is unclear what can be learnt from an unbiased estimate for a non-randomly selected subsample of the population of interest. Given the likely heterogeneity in impacts, the biased observational study for a random sample from the whole population may be closer to the truth.

Second, bias is not the only thing that matters. The appropriate decision rule for choosing an estimator (and designing a research study more generally) depends on the application. A popular statistical decision rule is to minimize the mean-squared error (MSE), i.e., the expected value of the squared deviation between the estimate and its true value. As is well-known in statistics, the MSE is the estimator's squared bias *plus* its variance.[19] Thus, this decision rule does *not* tell us that an unbiased estimator is always best.[20] MSE is not the only defensible decision rule—for example, one might ask how often the trials are within some absolute distance of the true value—but the point here is that unbiasedness is not all that matters.

The economics of impact evaluation comes into play here. Larger sample sizes reduce the variance of estimates. Many observational studies use existing data, from administrative records ("big-data"), as well as existing surveys. RCTs typically require new special-purpose surveys. Thus, for a given budget, RCTs will often have lower sample sizes and (hence) higher variances.

Nor is the outcome clear when a non-RCT requires new surveys. A good way to reduce bias is with better data. Longer survey questionnaires will probably entail smaller sample sizes for a given budget. But the data requirements for an RCT are unlikely to be different, noting that one wants baseline data to test for covariate balance in an RCT.[21] The additional randomization (for treatment) in an RCT is unlikely to be costless, and re-randomization may well be needed to

[19] If β is the true value and $\breve{\beta}$ its estimator then (by definition) $\text{MSE} \equiv ((\breve{\beta} - \beta)^2) = (E(\breve{\beta}) - \beta)^2 + E((\breve{\beta} - E(\breve{\beta}))^2)$. The first term on the RHS is the squared bias of $\breve{\beta}$ while the second term is its variance.

[20] This is well recognized in introductory econometrics texts. For example, Jonston (1984: 28) writes that "on the mean-squared-error criterion a biased estimator may be preferred to one with smaller or zero bias if its variance is sufficiently small to offset the larger bias." Also see the discussion in Green (1991: 97–9).

[21] Ex post balancing tests and retrospective adjustments are often recommended for RCTs (Cox and Reid 2000; Hinkelmann and Kempthorne 2008; Bruhn and McKenzie 2009; Hernán and Robins 2018).

assure covariate balance (Morgan and Rubin 2012). In medical applications, RCTs are widely thought to be more costly than observational studies.[22] I have not seen systematic cost data for development impact evaluations, though one often hears concerns about underpowered RCTs.[23] Cost comparisons for evaluations at the World Bank suggest higher costs for RCTs (though the comparisons are crude).[24] Development RCTs often encounter difficulties of implementation in the field that are not found for observational studies. To see what this can mean for the choice of method, suppose that each trial is drawn from one of two normal distributions, one for an RCT and one for a non-RCT. The parameters (its mean and variance) depend on the chosen method. The mean of the RCT distribution of trial results is taken to be the true mean, while it is not for the non-RCT. Even so, despite the bias, the variance of a non-RCT could be low enough to assure that it yields a higher share of its trials that are closer to the truth than the RCT. Figure 1.2 illustrates a hypothetical case, showing that even a biased observational study can be closer to the truth than an (unbiased) RCT.[25] Two densities are shown for impact estimates from both RCT and non-RCT designs, both drawn from normal

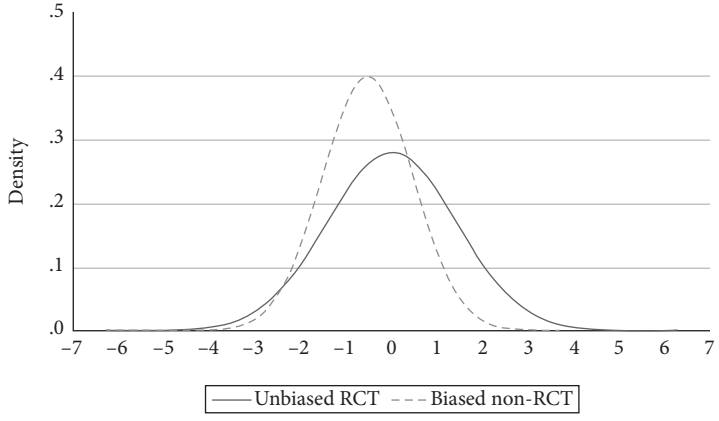

Figure 1.2 Density functions for the estimates of mean impact from two hypothetical designs for impact evaluations

Source: Author.

[22] See, for example, Hannan (2008) and Frieden (2017).

[23] For example, in reference to development RCTs, White (2014) says that "the actual power of many RCTs is only around 50 per cent. So, an RCT is no better than tossing a coin for correctly finding out if an intervention works." Sampling variability appears to account for half or more of the variability in impact estimates from RCTs; see Meager (2019), with reference to microcredit schemes.

[24] The World Bank's impact evaluations in recent times have tended to be RCTs with considerably higher average cost than the evaluations done in the International Financial Cooperation (within the World Bank Group), where observational studies are more common (World Bank 2012). This is at best suggestive since the comparison is not properly controlled.

[25] Green (1991, Section 4 0.3) uses a similar example to show that a more biased estimator can have lower MSE.

distributions. (The densities may or may not be conditional on covariates.) The true impact is zero, which is the mean of the distribution from which the RCT trials are drawn. The non-RCT trials are drawn instead from a distribution with a mean of –0.5, which is their systematic bias. The other difference is that the RCT trials are drawn from a distribution with a variance of 2, while for the observational study it is 1. This can be interpreted as saying that, for a given budget, the non-RCT allows double the sample size in each trial.

Which method does better, in that its trial estimates tend to be closer to the truth? Define "closer to the truth" as being more likely to be within a fixed interval centered on the true value—in this case, an interval $(-\delta, \delta)$ (for some $\delta>0$).[26] Figure 1.3 gives the percentage of trials close to the truth for each method. Suppose we define "closer to the truth" as an impact estimate in the interval (–0.5, 0.5). We find that the RCT gets an estimate that is within this interval for 27 percent of its trials, but this is so for 34 percent of the non-RCT trials. If instead we define "closer to the truth" as an estimate in the interval (–1, 1) then this is so for 52 percent of RCT trials versus 62 percent using the observational study. In this example, the observational study is closer to the truth for all δ!

Of course, this is only one of many possibilities, and one can readily construct examples where the RCT does better. Figures 1.2 and 1.3 only illustrate that the less biased impact evaluation need not get us closer to the truth. That is an open question as it depends crucially on the power of the trials that can be afforded given the budget. The key point is that we cannot rule out the possibility that, for

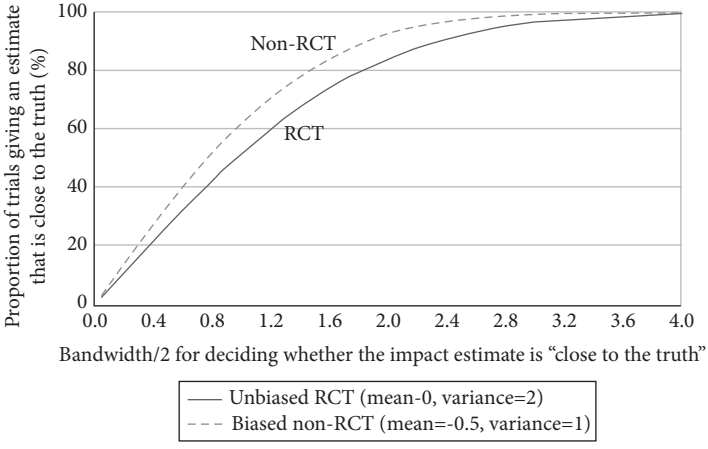

Figure 1.3 Proportion of trials giving an impact estimate that is close to the truth, comparing an unbiased RCT with a biased non-RCT on a larger sample

Source: Author.

[26] In some applications the interval need not be symmetric around the true value, i.e., errors in one direction are more costly to the agreed objective.

a given budget, the RCT ends up with less reliable impact estimates, that are often further from the truth than even a biased observational study. We lack a theoretical justification for the claimed (unconditional) "gold standard" hierarchy of methods.

In defense of RCTs, we currently know rather little about the distribution of the biases in observational studies, while (as noted) the unbiasedness of an RCT comes with an estimable variance, to facilitate calculation of its confidence interval. This points to the need for more research on the distribution of estimates from observational studies, such as by comparing estimates with those from RCTs for the same setting.[27] It remains, however, that if we insist on only doing RCTs (when feasible) then we may be forgoing observational studies that are more often close to the truth.

Greater clarity may well emerge when we know the context. If one knows the setting and program well enough to identify the relevant confounders—the model of how the program works—and can collect data on them, or find measurable deconfounders, then one may well obtain a very reliable impact estimate by observation alone. On the other hand, if there is little scope for collecting baseline data on the relevant confounders, and the unit cost of randomized assignment is not too high (so that reasonably large sample sizes are feasible with the available budget), then an RCT has much appeal. It is the widely heard "gold-standard" generalization that is at issue here.

We can go further and ask what design is optimal in the sense of minimizing (say) the MSE, while recognizing our uncertainty about the true model.[28] Let us assume that at least some of the baseline data are continuous covariates and that we have Bayesian priors on the model uncertainty. Then we can appeal to a result in Kasy (2016), namely that there exists a deterministic (non-random) assignment of treatment status based on the covariates that minimize the expected MSE.[29] Then there is no gain from randomizing the assignment given the covariates (and there will be efficiency gains from taking account of observables in RCTs). By implication, to justify a strong preference for an RCT one needs to attach some intrinsic value to randomization as an end in itself, and be willing to forgo accuracy in getting closer to the true mean impact. Advocates have sometimes fallen back on such methodological preferences, independently of precision in estimating impact; for example, Banerjee et al. (2017b) show that an RCT can

[27] See, for example, Chabé-Ferret (2018).

[28] Following Kasy (2016) this can be recognized as a problem in statistical decision theory, i.e., the choice of an estimation method to minimize a loss function based on the data actually available.

[29] This holds for any Bayesian risk function and for a minimax rule for the worst-case (Kasy 2016). Kasy provides software to implement the optimal assignment of treatment for minimizing the MSE. Note that a continuous covariate assures a unique optimum assignment of treatment status. With discrete covariates an RCT may do as well, but no better.

still dominate as long as one puts a high enough weight on the welfare of those who prefer RCTs.

The influence of the randomistas has stemmed in part from the (much-heard) belief that RCTs are the preferred statistical tool when feasible. This review of the underlying theory has cast considerable doubt on that belief. As we will see, other sources of their influence are no less questionable.

1.3 The Influence of the Randomistas on Development Research

Early examples of the use of RCTs in social policy contexts include the various experiments on US social policies starting in the 1960s.[30] With regard to the recent development applications, the 3ie database has 133 published RCTs over the period 1981–1999. The earliest RCT in the database is from a World Bank research project on education interventions (textbooks and radio lessons) to improve the math scores of students in Nicaragua, namely Jamison et al. (1981).[31] Among the pre-2000 RCTs, that done by the Government of Mexico for the *Progresa* evaluation, which started in 1997, is an especially notable example. The (generally positive) results in the literature generated by the data from that RCT were influential in the expansion of Conditional Cash Transfers to over 50 countries today.[32]

So, at the beginning of the new millennium, there was nothing new to the idea of RCTs in development applications. What changed is the popularity of that idea. The annual production of RCTs has been far higher since 2000 (Figure 1.1). Numerous individual academics and groups have contributed, but one group stands out, the Abdul Latif Jameel Poverty Action Lab (J-PAL).[33] This was founded in 2003 (as the Poverty Action Lab) and has been based in the Department of Economics at the Massachusetts Institute of Technology (MIT). The founders were Abhijit Banerjee, Esther Duflo, and Sendhil Mullainathan. At the time of writing, J-PAL's website[34] reports that they have 1012 completed and ongoing RCTs in 83 countries. For an academic research group to get that far in

[30] On the history of RCTs in US social policy see the discussions in Burtless (1995) and List and Rasul (2011). Other commentaries on the history of RCTs more generally can be found in Ziliak (2014) and Leigh (2018).

[31] Banerjee et al. (2019) claim that the use of RCTs in development was "kick-started" in 1994 in a study by one of the authors (Kremer). However, published development RCTs had existed for at least 13 years prior to that.

[32] On the *Progresa* impact evaluation and its influence see Skoufias and Parker (2001) and Fiszbein and Schady (2010).

[33] Another prominent group doing and promoting RCTs is the non-profit organization, *Innovations for Poverty Action* (IPA), founded in 2002 by Dean Karlan (then at Yale). IPA and J-PAL often work together, and clearly have close links. Within international organizations, the most prominent group doing RCTs is the *Development Impact Evaluation* (DIME) group at the World Bank; three-quarters of DIME's evaluations have used this method (World Bank 2016).

[34] https://www.povertyactionlab.org/

just 15 years is nothing short of amazing. On top of its own RCTs, J-PAL has clearly influenced the shift in emphasis in empirical development economics more broadly toward RCTs. Indeed, J-PAL's huge RCT output is unlikely to be even a majority of the total count of ongoing RCTs today.

The effort (centered on J-PAL but going further) came with a new prestige for this type of empirical research on development economics, as indicated by the award of the 2019 Sveriges Riksbank Prize in Economic Sciences (in Memory of Alfred Nobel) to Abhijit Banerjee, Esther Duflo, and Michael Kremer. As said in the headline of the announcement, the prize was awarded "for their experimental approach to alleviating poverty."

The discussion in this section looks first at the reasons for the influence of the randomistas on development research. It then asks if their influence has been justified.

Why have the randomistas had so much influence? Propagating the view that (when feasible) RCTs dominate purely observational studies in attempting to infer causation has clearly held sway. The landing page of J-PAL's website tells us that: "Our mission is to reduce poverty by ensuring that policy is informed by scientific evidence." Toward that aim, J-PAL only does RCTs. Strictly that does not imply that J-PAL's researchers think that observational studies are unscientific (and, independently of J-PAL, many J-PAL-affiliated researchers have used methods of observational study). However, in this context, the phrases "scientific evidence" and (another favorite, including on J-PAL's website) "rigorous evidence" are code for RCTs in the eyes of most readers, and that is plainly intentional. The implication is even stronger than the "gold standard" claim: for some advocates, RCTs are not just top of the menu of approved methods, nothing else is on the menu!

The appeal of RCTs reflects in part the challenges faced in identifying causal impacts. Since the 1990s, we have seen a welcome rise in the attention given to identification problems in economics,[35] though it has been argued that this partially displaced attention from other important issues, such as measurement errors (Gibson 2019). More critical attention has been given to the validity of IV estimators. It is easy to show that a failure of either of the aforementioned conditions for a valid IV can bias the estimate—possibly more so than for Ordinary Least Squares (OLS), which treats placement as exogenous. It was not hard for researchers to find exogenous variables that are correlated with selected treatment status (though they still needed to pass the appropriate tests). Accepting the exclusion restriction (that the IV is irrelevant to outcomes given treatment status and the control variables) on theoretical grounds was often far more challenging. There were some cases in which the IV could be readily accepted, but this was not always so. From the mid-1990s, seminar audiences and referees were regularly pointing to

[35] In common usage, a parameter is said to be "identified" when its value can, in principle, be derived from the data under certain assumptions.

reasons why the IV in specific papers could have an effect on outcomes that is independent of the endogenous variable. In due course, some economists started to reject almost any attempt at establishing causality without random assignment.

If one only wants to know the difference in mean outcomes between those assigned the option for treatment and those not—which is called the Intent-to-Treat (ITT) parameter—then randomization side-steps these concerns about IV estimates. Given randomization, the treatment assignment is exogenous, uncorrelated with the regression error term. However, ITT can be a rather limited parameter. It is sometimes defended as "policy-relevant" in that the policy is often the assignment of the *option* for treatment. Yet how would you react to finding that the mean impact is (say) zero among those offered treatment, but positive among those who took it up? Such a finding would surely be of interest to policymakers and citizens. In learning from an RCT, a prospective adopter of the treatment will want to know mean impact for those treated.

Take-up of a randomly assigned treatment with human subjects is never assured, and compliance is typically endogenous. So the econometric problem often returns in practice. The randomistas have a solution: use randomized assignment as the IV for actual treatment. Clearly, take-up requires assignment, so this IV is correlated with treatment status. Since it is random, the IV is also uncorrelated (in expectation) with the error term when the treatment effect is common across the population. (The discussion will return to the complications that can arise when impacts vary in a way that is unknown to the researcher but known to each participant, who responds accordingly.)

Beyond these econometric arguments, a number of other factors contributed to the randomistas' growing influence from the early 2000s. Researchers who did not use randomization started to be criticized by the randomistas, and their papers started to get ignored in citations to the relevant literature. Some of this took the form of referees' comments on journal articles, which are not public. Journal editors do not need to accept such critiques, though the leading randomistas appear to have been influential and in due course became prominent among the editors and editorial boards of economics journals. At times, the critiques also took a public form, such as the study by Finkelstein and Taubman (2015), which questioned the fact that observational and other non-random methods are often used in evaluating health-care delivery policies. This finding was then reported in the *New York Times* under the heading "Few Health System Studies use *Top* Method, Report Says" (Tavernise 2015; my emphasis), where the "top" is explicitly taken to be an RCT. The message here is clear, though it is less clear that it is right. Some public health specialists have argued that there has been too much attention to evaluations for individual treatments at the expense of research on health systems.[36]

[36] See, for example, Rutter et al. (2017).

The leading randomistas also did a good job in teaching others how to use their preferred method.[37] Development economists got up to speed quickly, as did some NGOs. They have also been steadily raising the bar on what constitutes a good RCT, though the observation of Heckman and Smith (1995) that RCTs get less critical scrutiny than other methods still seems true today.

Another factor enhancing their influence is that J-PAL's founders professed their desire to make the world a better place through evidence-based policy-making. This was J-PAL's declared motivation from the outset. By this view, doing many RCTs lets us figure out what works and what does not, to scale up the former and scale down the latter (Banerjee 2006). An analogy is drawn with RCT's in clinical trials, as used to find out what drug works best on average (Favereau 2016).

Some followers have clearly been attracted by the zeal of the leading randomistas. By this view, "the experimental ethic has been proposed as the way to change the spirit of development" (Donovan 2018: 27). The randomistas can be seen in part as an epistemic movement that attracts its "true believers"[38] who advocate RCTs with "near-religious zeal" (Heckman, Chapter 12, this volume). The movement's faith in RCTs promises its followers a "quiet revolution" (Banerjee and Duflo 2011: 265).

Supporters (including donors) have also been attracted by the simplicity of RCTs—that they are "more transparent and easier to explain" (Duflo 2017: 17). It is easier for non-economists to understand an RCT than the methods often favored in observational studies, which were also getting increasingly sophisticated, and technically demanding, by the time J-PAL was founded.

Is the randomistas' influence justified? As Section 1.2 argued, the statistical foundations do not tell us that (when feasible) RCTs are invariably more reliable, whatever the context, and so sit at the top of the hierarchy of methods. This is more a matter of faith than science. The rejection of methods using non-random assignment in some quarters has clearly been an over-reaction to the challenges faced in identifying causal effects this way.

Nor is the analogy to clinical trials persuasive. It is unclear that the idea of using black-box RCTs to figure out what works and what does not in development is feasible given the dimensionality in both interventions and contexts. Too often, the arguments made for RCTs lack a clear economic rationale for the intervention, or a coherent structure for understanding why it may or may not work (Heckman and Smith 1995).

While the development randomistas were pointing to clinical trials as the model, medical researchers were taking a more nuanced view.[39] On the one hand,

[37] An example is the excellent "RCT toolkit" produced by Duflo, Glennerster, and Kremer (2011). The World Bank's Development Impact blog has provided a great deal of useful methodological support for doing RCTs.

[38] A reviewer of Leigh (2018) describes the author as a "true believer" and then recounts the various personal choices that Leigh makes in his life based on the results of RCTs (Wydick 2018).

[39] Examples of the following points are found in Concato, Shah, and Horwitz (2000), Silverman (2009), Bothwell et al. (2016), and Frieden (2017).

some of the recent literature suggests that past concerns about bias in causal observational health and medical studies have been exaggerated. On the other, it now seems to be accepted that gains from removing systematic bias need to be weighed against the costs and risks of clinical RCTs.

Yet, putting these points to one side, it must be recognized that the medical context is different. Economists (and other social scientists) are dealing with people (as individuals and groups) in social and/or economic contexts in which they can be expected to exhibit greater heterogeneity, and almost certainly greater agency, than is likely in clinical trials. We may often know rather little about the specific setting a priori.

Some deeper inferential issues lie under the surface of the randomistas' claims—issues that are known to the experts on both sides but poorly understood more broadly. There is almost certainly some unobserved heterogeneity in the impacts of treatment. There are many sources, including both the circumstances of the individual (such as past experience with the type of intervention) and the effort made by agents (reflecting their beliefs about the impact).[40] Such heterogeneity raises the question of "impact for whom?" This was answered by Angrist, Imbens, and Rubin (1996), who showed that the IV estimator is giving the mean impact for a subset of the treated, namely the "compliers," induced to switch their treatment status by the randomized assignment.[41]

When estimating the mean impact on those treated, the validity of randomized assignment as the IV to address selective take-up can be questioned in the presence of behavioral responses to such unobserved heterogeneity in the impacts of treatment (Heckman and Vytlacil 2005; Heckman Urzua, and Vytlacil 2006).

The differing impacts must then be relegated to the regression error term, interacting with the selective take-up of the randomized assignment. Those units with high returns to treatment will be more likely to take it up. Then the interaction effect that has now surfaced in the error term must be correlated with the randomized assignment. The exclusion restriction fails. (Of course, this does not matter if one only wants ITT.)

Identifying the impacts of social programs is rarely easy, with or without randomized assignment. Suppose that the latent characteristics that enhance impact at the individual level also matter to the counterfactual outcomes in an RCT with selective compliance. The choice of estimation method then depends crucially on what impact parameter one is interested in, the type of program one is evaluating and the behavioral responses to that program (as shown in Ravallion 2014). If the latent factors leading to higher returns to treatment are associated with lower counterfactual outcomes then the "IV cure" for endogenous treatment can be

[40] On the latter source see Chassang et al. (2012), who study the implications for the external validity of RCTs.

[41] Also see the discussion in Pearl (2009, Chapter 8).

worse than the disease. Indeed, the OLS estimator may even be unbiased, despite the selective take-up. The key point is that practitioners need to think carefully about the likely behavioral responses to heterogeneous impacts in each application—similarly to any observational study.

The design of RCTs in practice can also pose threats to identification. The randomized assignment is sometimes done across clusters of individuals, such as villages. Some clusters get the treatment and some do not. Those within a selected treatment cluster are left free to take up the treatment as they see fit. This is a now classic design in development applications.[42] It runs into a problem whenever there is interference within the clusters whereby non-participants in the selected treatment clusters are impacted by the program. For example, the cluster RCT in Ravallion et al. (2015) used an entertaining movie to teach people their rights under India's National Rural Employment Guarantee Act. It was impossible to enforce ticket assignments within villages; the movie had to be shown in public places—often open areas of the village. So access to the movie was randomly assigned across villages, with people free to choose whether to watch it. Some did not, but (of course) they can talk with others who did, and this turned out to be an important channel of impact on knowledge. The cluster randomization had to be combined with a behavioral model of why some people watched the movie (Alik-Lagrange and Ravallion 2019). Only then could the direct treatment effect (watching the movie) be isolated from the indirect effect (living in a village with access to the movie). In this example, the spillover effects within clusters violate the exclusion restriction, so the use of cluster assignment as the IV for individual take-up performs poorly.

The generic point is that—contrary to the claims about clean identification of the mean causal impact using randomized assignment—assumptions and models are often required in practice. It does not help that the behavioral assumptions underlying studies using randomization are not always explicit (Keane, 2010). Structural approaches, in contrast, force this to happen.

Some concerns have received less attention in the literature than they merit. One example is Hawthorne effects, whereby monitoring changes behavior. (For example, if you know you are in the control group you may be inclined to seek a substitutable treatment. Or some in the treatment group may try to please the experimenter.[43]) RCTs in economics do not often have the double-blind feature common to clinical trials, so biases associated with monitoring are more likely,

[42] Of course, if one can use double randomization—randomizing within villages as well as between them—then one can readily address this type of interference (Baird et al., 2017). Cluster randomizations are designed for situations in which within-cluster randomization if not feasible. Such situations are common in development applications.

[43] One RCT randomly assigned knowledge about the experimenter's intent, but did not find any significant effect (Mummolo and Peterson 2019). Further tests of this sort are needed.

and they merit more attention in development applications.[44] A second example is the topic of the next section.

1.4 Taking Ethical Objections Seriously

Ethical concerns are never far removed from policy-making. There are two dangers of not taking the ethics of evaluation seriously. First, morally unacceptable evaluations may end up being done, and possibly more often in poor places with vulnerable populations and weak institutions for protecting their rights. Second, socially valuable evaluations may be blocked as too risky politically, largely in ignorance of the benefits.

RCTs have been criticized on the grounds that "randomizers are willing to sacrifice the well-being of study participants in order to 'learn'" (Ziliak and Teather-Posadas 2016).[45] Critics have often pointed out that in an RCT some people who need the treatment are not getting it, while others receive a treatment they do not need. The criticism is also heard that RCTs in poor countries do not get the same ethical scrutiny that is expected (though by no means assured) in rich countries.[46] In using RCTs for clinical trials of potentially hazardous treatments, there have been some well-documented cases in which participants in developing countries were largely unaware of the health risks they faced if they end up being treated.[47] Baele (2013) argues that the development randomistas have not paid enough attention to the ethics of their RCTs. Glennerster and Powers (2016) offer a cautious ethical defense of RCTs against their critics.

Ethical validity is not a serious issue for all evaluations. Sometimes an impact evaluation is built onto an existing program such that nothing changes about how the program works. The evaluation takes as given the way the program assigns its benefits. So if the program is deemed to be ethically acceptable then this can be presumed to hold for the evaluation. We can dub these "ethically benign evaluations."

Other impact evaluations deliberately alter the program's (known or likely) assignment mechanism—who gets the program and who does not. Then the ethical acceptability of the intervention, as it normally works at scale, does not imply that the evaluation is ethically acceptable. Call these "ethically contestable evaluations." The main examples in practice are RCTs. Scaled-up programs

[44] This aspect of the difference between economic RCTs and clinical RCTs is discussed further in Favereau (2016). For a useful overview of the Hawthorne effect in the health field see Friedman and Gokul (2014).

[45] Also see the comments in Barrett and Carter (2010), Baele (2013), and Mulligan (2014).

[46] In the US, the ethics of using RCTs for the evaluation of Federal social policies has not received the same attention as for clinical trials. Blustein (2005) discusses the reasons.

[47] See, for example, Sathyamala (2019) on an RCT used to study the health risks of a contraceptive drug in Africa.

almost never use randomized assignment, so the RCT has a different assignment mechanism, with potentially large differences in the benefits, given the likely heterogeneity in impacts. An RCT can be contested ethically even when the real program is fine.

It is surely a rather extreme position (not often associated with economists) to say that good ends can never justify bad means. It *is* ethically defensible to judge processes in part by their outcomes; indeed, there is a long and respected view in moral philosophy that consequences often trump processes—with utilitarianism as the leading example. It is not inherently "unethical" to do an RCT as long as this is deemed to be justified by the expected benefits from new knowledge. However, the consequential benefits do need to be carefully weighed against the process concerns. This is especially so in the (many) instances in which a feasible, and ethically benign, observational study is an option.

Ethics has been much discussed in medical research where the principle of equipoise requires that there should be no decisive prior case for believing that the treatment has impact.[48] Only if we are sufficiently ignorant about whether it is better to be in the treatment group or the control should we randomize at all, or continue with an RCT.[49] If evaluators are to take ethical validity seriously then some development RCTs will have to be ruled out as unacceptable, given that we are already reasonably confident of the outcomes—that the gain from knowledge is not likely to be large enough to justify the ethically-contestable research.[50]

The principle of equipoise is rarely applied to RCTs for development and social policies. Indeed, there may well be a tendency in the opposite direction. A recent call-for-proposals from a prominent philanthropic funder gave explicit preference to any RCT proposal "That is backed by highly-promising prior evidence, suggesting it could produce sizable impacts on outcomes…" (Laura and John Arnold Foundation 2018: 2). At one level, one can understand the funder's preference, given that RCTs are costly and there is a desire to have impact with limited resources. Some ex ante filters of this sort make sense. (One would not want to fund an RCT for an intervention that is unlikely to turn out to be feasible on the ground.) However, the above example points to a tension between donor objectives and ethical concerns. Ex ante confidence of "sizeable impacts on outcomes" leaves one worried about withholding a treatment from those who need it (and wasting treatment on those who do not). This also points to a concern about the funding processes determining what gets evaluated. Section 1.5 returns to this topic.

[48] There is a good discussion in Freedman (1987), which introduced the principle of equipoise in clinical trials. In the context of development impact evaluations, see Baele (2013) and McKenzie (2013).

[49] The "we" here is best thought of as a set of people with sound knowledge of the relevant literature and experience. This is sometimes called "community equipoise."

[50] See the examples discussed in Barrett and Carter (2010), Ziliak and Teather-Posadas (2016) and Narita (2018).

There have been some ethical defenses of RCTs. One view is that RCTs are justified whenever rationing is required; when there is not enough money to cover everyone, it is argued that randomized assignment is a fair solution.[51] This makes sense when information is very poor. In some development applications, we may know very little ex ante about how best to assign participation to maximize impact. Nevertheless, when alternative allocations are feasible *and* one does have prior information about who is likely to benefit, it is surely fairer to use that information, and not randomize, at least unconditionally.

It has also been argued that the method of conditional randomization (also called "blocked" or "stratified" randomization) can relieve ethical concerns. The idea here is that one first selects eligible types of participants based on prior knowledge about likely gains, and only then randomly assigns the intervention, given that not all can be covered. For example, if one is evaluating a training program or a program that requires skills for maximum impact, one would reasonably assume (backed up by some evidence) that prior education and/or experience would enhance impact, and then design the evaluation accordingly. This has ethical advantages over pure randomization when there are priors about likely impacts.

There is a catch. The set of things observable to the evaluator is typically only a subset of what is seen on the ground. At (say) village level, there will often be more information than is available to the evaluator—information revealing locally that the program is being assigned to some who do not need it, and withheld from some who do. But whose information should decide the matter? Pleading ignorance seems a lame excuse for an evaluator when other stakeholders do in fact know very well who is in need and who is not.

It has also been argued that encouragement designs are less contentious ethically. The idea is that nobody is prevented from accessing the primary service of interest but the experiment instead randomizes access to some form of incentive or information. This does not remove the ethical concern—it merely displaces it from the primary service of interest to another space. Ethical validity still looms as a concern when the encouragement is being deliberately withheld from some people who would benefit and given to some who would not.

Consider, for example, the RCT in Bertrand et al. (2007). One treatment arm provided a large financial reward to those participants who could quickly obtain a driver's license in Delhi India, which facilitated bribes to licensing officials. The RCT did not pay bribes directly or give out licenses to people who did not verifiably know how to drive, but these were predictable outcomes. The expected gain from this RCT was a clean verification of the proposition that corruption happens

[51] See, for example, Goldberg's (2014) comments on Mulligan (2014). The same point is made by Fiennes (2018).

in India and has real effects. However, there does not seem to have been any serious prior doubt about the truth of that claim.

RCTs can be designed to help address ethical concerns. One option is to use an "equivalence trial" for which the control group gets what is thought to be the next best treatment.[52] Possibly in contrast to biomedical settings, there may be little agreement on the *best* option in each specific development application. Nonetheless, it seems unlikely that the common use of the "do-nothing" or placebo control would pass close ethical scrutiny in most development applications. There is usually some option. (Nor is "doing nothing" likely to be a particularly relevant counterfactual for most policy-makers.)

Another option is adaptive randomization. This is feasible when there is a sequencing of assignment, with observed responses at each step. Adaptive randomizations change the assignment along the way, in the light of the accumulated evidence on impacts.[53] Narita (2018) has proposed an interesting market-like adaptive design for social experiments, whereby one takes account of each participant's willingness-to-pay for the chance of treatment, given prior knowledge about impacts.[54] Unlike a classical RCT, one ends up with a Pareto efficient experiment, though with similar statistical properties for the impact estimates. At the time of writing, this idea does not appear to have been implemented in the field.

In the US and elsewhere, Institutional Review Boards (IRBs) have become common for proposed studies with human subjects. There is a designated IRB for most research institutions. They are largely self-regulating. Beyond occasional anecdotes, there does not appear to have been a systematic assessment of how well IRB processes have worked for development RCTs. One thing seems clear: IRBs need to give more attention to assessing the expected benefits of an ethically contestable evaluation given prevailing knowledge. Syntheses of current knowledge can help and these are becoming more common.[55]

If pressed, many randomistas acknowledge the ethical concerns reviewed above, though they rarely give them more than scant attention in their papers. They *assume* (more often implicitly) that their RCTs generate benefits that outweigh such concerns. Whether that is true is rarely obvious, and merits more attention.

We should also ask how well research efforts match the knowledge gaps. Imbalances of this sort raise further ethical concerns, given pressing development challenges and limited resources for research. The next section takes up these issues.

[52] This idea has been a much debated in biomedical applications, notably in the context of the revisions done in 2000 to the World Medical Association's 1964 Helsinki Declaration. For further discussion see Levine (2006).

[53] These are getting serious attention in biomedical research. For example, the US Food and Drug Administration (2010) has issued guidelines for adaptive evaluations. Also see Cox and Reid (2000, Chapter 3).

[54] Also see Chassang et al. (2012) and the discussion in Özler (2018).

[55] These are sometimes referred to as systematic reviews; see for example, the 3ie searchable database and the Campbell Collaboration on such reviews.

1.5 Relevance to Policy-making

While there is clearly a lot more to good policy-making than good evidence, policy-makers increasingly turn to evidence, hoping to inform their choices, and win political debates. The policy-relevance of evaluative research matters.

To my knowledge, there has not yet been a comprehensive and objective assessment of the influence on development policy of all those RCTs. Nonetheless, one can point to examples of policy-relevant research using RCTs. To give just one example, Banerjee et al. (2015a) used RCTs in six countries (Ethiopia, Ghana, Honduras, India, Pakistan, Peru) to evaluate the long-established approach taken against poverty by BRAC using a combination of transfers (assets and cash) targeted to the poorest with literacy and skill training.[56] The researchers found economic gains from adopting BRAC's approach some three years after the initial asset transfer, and one year after the disbursements finished. If one is willing to extrapolate the earnings gains into the distant future—although that is clearly a strong assumption—then their present value often exceeds the cost of the BRAC-type program (Banerjee et al. 2015a).

Without aiming to provide a comprehensive assessment, this discussion points to some limitations of RCTs for informing development policy, drawing on the literature.

Policy-relevant parameters: Even under ideal conditions, an RCT is only well-suited to estimating a rather narrow subset of the parameters of interest to policy-makers. In reality, one expects that there will be both gainers and losers, depending on the context and the characteristics of participating units (and, as noted, some of those characteristics are unobserved to the analyst, though still motivators of behavior, including whether or not to take up the treatment). There is a distribution of impacts. Policy-makers may want to know what proportion of the population benefit, and what proportion lose, or what types of people gain and what types lose. Identifying these policy-relevant parameters will typically require more data and more structural-econometric methods. A full-blown structural model need not be essential for addressing the question of interest, but (at the other extreme) an RCT will rarely deliver what is needed.

There are ways of reliably learning more about individual impacts than simply their mean. For example, the Local Instrumental Variables estimator proposed by Heckman, Urzua, and Vytlacil (2006) aims to identify the marginal treatment effects (MTEs) at all values of the empirical probability of being treated. Unlike a standard RCT, "selective trials" allow one to identify the MTEs by basing the probability of assignment to treatment (rather than control) on agents' expressed

[56] BRAC now stands for Building Resources Across Communities. The NGO started in Bangladesh (where it was once called the Bangladesh Rural Advancement Committee) but now works in many countries.

willingness to pay (Chassang et al. 2012). One can then aggregate up to get the mean impact, as would be identified by an RCT. But one learns a lot more than the average impact.

Sometimes it is also possible to reliably ask counterfactual questions in surveys. This is done in Murgai, Ravallion, and vand de Walle (2015), who asked participants in a workfare program what they think they would be earning otherwise (with observational checks against local labor markets). Then one can learn more about the distribution of impacts, though (of course) there are measurement errors in survey responses, so some averaging will almost certainly be required.

An aspect of performance that is often of interest to policy-makers is who benefits from the program, as determined (in part) by the assignment mechanism implied by its design. If it is demand driven, what are the characteristics of those choosing to take it up? If it is rationed, to whom? Such questions come at the first stage in an important class of observational methods using matching that start with a statistical model of who gets the program and who does not.[57] Of course, if it is an RCT then, in expectation, the assignment is not predictable, and if there is full compliance then nothing can be learnt about the types of people likely to participate when the program is scaled up.

With imperfect compliance, we can learn about this from the first stage of the aforementioned IV estimator. Indeed, as Heckman and Pinto (2019) argue, once we recognize that take-up of the randomized assignment is the outcome of rational choice, we can use it to study both the determinants of participation and identify a wider range of casual parameters. For example, by varying the incentives in the classic RCT, and invoking the weak axiom of revealed preference, the results in Heckman and Pinto (2019) can be applied to the problem of low take-up of social policies, which is often common among poor and/or socially excluded people. Instead of thinking about selective compliance by human subjects as a statistical nuisance we can learn from it.

An RCT might also be used to assess likely impact ex ante, and then later do a separate evaluation of the actual program at scale, using an observational estimator. This sounds promising but it should be understood that, given selective take-up and heterogeneous impacts, one has essentially evaluated two different programs, only one of which is actually implemented by the government. It is not hard to guess which will be of greater interest to policy-makers. Will the second evaluation be done? Possibly not if one takes the "gold standard" view.

At the heart of the problem of learning about policy effectiveness is that an RCT is a rather artificial construction, unlike almost any imaginable real-world policy.

[57] This refers to propensity-score matching. The predicted values of that model are the "propensity scores" used in selecting observationally balanced treatment and comparison groups (Rosenbaum and Rubin 1983).

External validity: Policy-makers naturally want to learn from such an experimental trial about how the same intervention might perform in another setting. This is a question about external validity. This can be in doubt for a number of reasons, including monitoring effects, general equilibrium effects, sampling problems and specific care in providing the treatment in the RCT (Duflo, Glennerster, and Kremer 2011).

Such issues are often ignored in papers documenting development RCTs, or the issues are only given a superficial treatment. For the majority of the 54 development RCTs published in eight economics journals (2009–14), Peters, Langbein, and Roberts (2018) find that the sources of external invalidity are not addressed and the information to address them is not provided. If different RCTs on a given intervention tended to agree then we can be more confident about external validity. But that is not the case. Vivalt (forthcoming) has documented the variance found in the impact estimates for a given program across settings (and even types of evaluators). Her findings warn against generalizations. As Vivalt also notes, poor documentation of contextual factors does not help. Pritchett and Sandefur (2015) provide examples (for microcredit schemes) in which a (presumed) internally valid RCT done in one context is inferior to an observational study for predicting impact in another context. Not all of this variability in estimates is due to heterogeneity in the true impacts; an estimate for seven microcredit RCTs found that 60 percent of the variability is due to sampling variation (Meager 2019). In practice, policy-makers will not be able to easily distinguish sampling variation from true impact variability.

The advantages of working with NGOs in doing RCTs (Section 1.2) have also raised questions about external validity. An example is found in the RCT on schooling in Kenya by Duflo, Dupas, and Kremer (2015). Randomly chosen schools were given the resources to hire an extra teacher working on a short-term contract. Children with the contract teachers were found to do significantly better in test scores than those with regular civil-service teachers. This experiment was implemented by a local NGO. However, Bold et al. (2018) attempted to replicate this at scale, using a follow-up RCT, but this time with an arm implemented by the government (as well as one by the NGO). This revealed that it was NGO-implementation that led to test-score gains, not the type of teacher. The teacher-effect found by Duflo, Dupas, and Kremer (2015) had vanished.

A "black box" reduced-form estimate (whether from an RCT or not) is not very informative for many purposes of policy-making. Learning from RCTs poses specific problems. Consider how we might learn about scaling up from an RCT (which is surely an important aim). An RCT randomly mixes low-impact people (for whom the program has little expected benefit) with high-impact people, based on latent attributes. It is plausible that the scaled-up program will have higher representation from the high-impact types, who will be attracted to the

program.[58] Given this purposive selection based on the (heterogeneous) expected impacts, the national program is fundamentally different to the RCT, which may contain little useful information for assessing the program at scale.

This reflects a more general point made by Moffitt (2006) that many things can change—inputs and even the program itself—on scaling up a pilot. An NGO keen to demonstrate its worthiness to attract funders will have an incentive to show impact from a trial that is not typical of its normal operations. Young researchers doing a field trial may apply greater effort than the government officials implementing the scaled-up version. External validity imposes constraints on the design and execution of pilots that are not given sufficient attention in practice.

One approach to learning about external validity is to repeat the evaluation in different contexts. For example, using an observational method, Galasso and Ravallion (2005) studied the performance of Bangladesh's Food-for-Education program in each of 100 villages and correlated the results with characteristics of those villages. The differences in performance were partly explicable in terms of observable village characteristics, such as intra-village land inequality (with more unequal villages being less effective in reaching their poor). Not allowing for such differences has been seen as a serious weakness in past evaluations.[59] Looking inside the black box of an impact evaluation can throw useful light on its external validity and policy implications. This will often require information external to the original evaluation design. An example is the *Proempleo* RCT by Galasso, Ravallion. and Salvia (2004). Vouchers for a wage subsidy were randomly assigned across people currently in a workfare program, with a randomized control group. The theory is that the wage subsidy will reduce labor costs to the firm and so make hiring the worker more attractive. Consistently with the predictions of the theory, the RCT found a significant impact on employment. However, subsequent checks against administrative records revealed a very low take-up of the wage subsidy by firms. So *Proempleo* did not work the way the theory had assumed. Follow-up qualitative interviews with firms and workers indicated that the vouchers had credential value to workers—a "letter of introduction" that few people had (and the fact that it was allocated randomly was a secret locally in this RCT). This could not be known from the RCT, but required supplementary observational data. (And this had not been anticipated by the researchers ex ante, so rigid adherence to a pre-analysis plan would have missed a crucial, policy-relevant, aspect of why the program had impact.) The extra data also revealed the importance of providing information about how to get a job, which carried implications for scaling up. However, scaling up the wage subsidy based on the RCT would have been a mistake.

[58] This is an instance of what Heckman and Smith (1995) dubbed "randomization bias." Also see the discussion in Heckman (Chapter 12, this volume), revisiting this issue in the wake of the boom in development RCTs.

[59] See for example the comments by Moffitt (2004) on trials of welfare reforms in the US.

A strand of the literature used randomization (either of the intervention or of some key determinant of its placement) to throw light on deeper structural parameters. This was done in some of the earlier applications to social policy evaluation in the US (Heckman, Chapter 12, this volume). In an example from recent development applications, Todd and Wolpin (2006) use the aforementioned RCT for *Progresa* in Mexico to model the dynamic behavioral responses to the schooling incentive provided by that scheme. Such research can help us understand a program's impacts and facilitate simulations of alternative policy designs. Todd and Wolpin show that a switch of the *Progresa* subsidy to higher levels of schooling would enhance overall impacts. In a similar vein, there is scope for using an RCT to test one or more key links in the "theory of change" underlying a program's rationale, even if the tool is not applicable to the program itself. This echoes the arguments of Heckman (1992) and Heckman and Pinto (2019) on the scope for more ambitious experiments informed by theory.

Knowledge gaps: To help antipoverty policy-making, researchers should ideally be filling the gaps between what we know about the effectiveness of policies and what policy-makers need to know. This is clearly not happening as well as we might hope. For example, Kapur (2018) recounts an interview with a former Chief Economic Advisor of the Government of India (GOI): "When asked how many of these expensive RCTs had moved the policy needle in India, Arvind Subramanian, Chief Economic Advisor, GOI, was hard pressed to find a single one that had been helpful to him in addressing the dozens of pressing policy questions that came across his table."[60]

Why do these knowledge gaps exist? There are random factors but there are also more systematic "knowledge market failures" (Ravallion 2009b). One source is the existence of externalities in evaluations. There is evidence that having an impact evaluation in place for an ongoing development project can help improve some aspects of its implementation, such as its speed of disbursement (Legovini, Di Maro, and Piza 2015). However, the knowledge gains from an evaluation also bring benefits to future projects, which (hopefully) draw on the lessons learnt from prior evaluations. Current project managers cannot be expected to take proper account of these external benefits when deciding how much to spend on evaluating their own project. There are clearly larger externalities for some types of evaluations, such as those that are more innovative—the first of their kind. The externalities in evaluation also play a role in the "myopia bias" that has been noted in development applications, such that long-term evaluations are rare (Ravallion 2009b; Bouguen et al. 2019).

Knowledge market failures also stem from publication biases originating in both the selection processes of journal editors and the behavior of authors,

[60] Also see the comments in Basu (2014) (another ex Chief Economic Advisor of GOI.)

including in documentation. Null results are less likely to be published or even written up.[61] Subsequent replications of experiments in economics often find less strong effects.[62] In some cases, the prior results have been adequately replicated but in the process have been found to be highly sensitive to questionable aspects of the data analysis that had not been obvious in the original paper.[63]

The dynamics of publication processes are a further source of persistence in knowledge gaps. Errors occur in the literature and it can take time to correct them. In recognition of its originality, the first paper on a topic may well be published prominently. Subsequent papers will tend to be relegated to lesser journals, cited less often, or may even have a hard time being published at all. The author of the original paper becomes the gatekeeper of knowledge on the topic. The gatekeeper is sometimes passable, but still has considerable influence. However, the first paper may not have got it right. On top of this, the incentives for effort at replication appear to be weak in economics.[64] (Yet in the sciences, failures to replicate have been common; see Ioannidis, 2005a.) Thus, the first draw from the distribution of impacts can have a lasting distortionary effect on accepted knowledge.

External invalidity also raises concerns about the process of knowledge accumulation. Even if the first paper came close to the truth in the specific context, it may have limited validity in different circumstances. When the topic concerns the impact of a policy, or an issue that is very relevant to that impact, policy knowledge will tend to be skewed accordingly.

These are generic concerns, not confined to RCTs. However, the "gold standard" method-hierarchy could well make things worse, as we will now see.

Matching research efforts with policy challenges: Knowledge gaps also stem from misalignments of evaluative effort. One aspect is that development evaluators too often ignore the scope for *fungibility*. Recipients (governmental or not) can re-allocate their own efforts in response to new funds, such as development aid. Donors are often implicitly funding something else. A less well-known implication is that donors and higher levels of government may well be evaluating the

[61] Among 221 social science studies it was found that "Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up" Franco, Malhotra, and Simonovits 2014: 1502). The distribution of reported p-values in papers published in the AER, QJE and *Journal of Political Economy* suggests that researchers tend to make specification choices that inflate the significance of their results to get over a "5% significance" hurdle (Brodear et al. 2016). Christensen and Miguel (2018) survey the evidence on these biases in published economic research and discuss how the biases might be reduced.

[62] Camerer et al. (2016) replicated 18 laboratory experiments published in the *American Economic Review* (AER) and *Quarterly Journal of Economics* (QJE). On average, the replicated effect size was one third lower than the original.

[63] See, for example, Bédécarrats et al. (2019a), which casts doubt on both the internal and external validity of the original RCT by Crépon et al. (2015).

[64] See the discussion in Rodrik (2009). Since then, 3ie has supported replication efforts for development impact evaluations through its Replication Window and its *Journal of Development Effectiveness*.

wrong thing from the point of view of assessing their own impact—they evaluate the project that the aid recipient put up for funding rather than the project that was actually funded, given the scope for fungibility. Then evaluative efforts are misaligned with development efforts.

RCTs are not to be blamed for this. However, strong methodological preferences on the part of evaluators can readily reinforce the misalignment. The development randomistas have had both output and substitution effects on knowledge. There is at least the suggestion of a positive output effect in the fact that we have seen a great many more RCTs since 2000 (Figure 1.1). However, as discussed already, neither the internal nor the external validity of these development RCTs is fully evident. We do not know the counterfactual—what we would have learnt if those resources (financial and human capital) had been deployed elsewhere.

The substitution effect relates to the methods used. Take, for example, the World Bank. While the earliest RCT in the 3ie database is by the Bank, until the early 2000s the tool was seen as only one of many credible options for IE. Since then there has been a marked switch in favor of RCTs within the Bank, which has been applauded by some observers; for example, an editorial in *The Lancet* declared (in ignorance of more than the history) that "The World Bank is finally embracing science." (Lancet 2004: 731).[65] The Bank's Independent Evaluation Group (IEG) reports that over 80 percent of the impact evaluations starting in 2007–10 used randomization, as compared to 57 percent in 2005–06 and only 19 percent in prior years (World Bank 2012).

Even if we presume that all those RCTs had a positive output effect on knowledge, the substitution effect could well work in the opposite direction. There are three aspects of the substitution effect. First, the emphasis on identifying causal impacts using RCTs has deflected attention from other methods of empirical investigation, including descriptive research, which is surely undervalued in development research today. Some of the policy lessons emerging from RCT research papers could have been derived from good "thick" descriptions (using qualitative and/or quantitative methods) of the real-world processes linking interventions to outcomes.

Second, there is a concern that the emphasis on assigned individualized programs has deflected attention from systemic research, typically using structural models. In economics more broadly, the decline in attention to structural work in teaching and research has been noted by Keane (2010) and others. This has also been a raised as a specific concern for research on public health (Rutter et al. 2017).

Third, a problem in evaluating the impact of the *portfolio* of development policies is that randomization is only feasible for a non-random subset of policies

[65] On the influence of RCTs at the World Bank see Webber and Prouse (2018).

and settings. The implication is that we lose our ability to make inferences about a broad range of policies if we rely solely on RCTs. As a generalization, randomization tends to be better suited to programs with clearly identified participants and non-participants, relatively short time horizons, that do not require imposing charges/taxes, and for which there is little scope for the costs or benefits to spill-over to the group of non-participants. Thus RCTs make more sense for private goods, which are easy to assign across individual households, than public goods with benefits shared across many people (Hammer 2017). There are exceptions (such as certain local public goods). However, it is generally far more difficult to randomize the location of medium- to large-scale infrastructure projects and seemingly impossible to randomize sectoral and economy-wide reforms. This makes the tool of limited use for some core activities in any country's development strategy.

Evaluations of the impacts of providing private goods beg for an economic rationale for the "policy." Would not markets provide the private good efficiently, eliminating the need for any impact evaluation? There may be good reasons why an evaluation for a private good is needed in specific contexts, but more often it seems that the randomistas are simply chasing opportunities for randomization. Granted, redistributive goals are mentioned at times, but in a rather casual way. Distributional impacts (such as on poverty) are rarely addressed with any rigor, or even identified as explicit outcomes. In short, the public economics is often missing.

To give an example of how an insistence on using RCTs distorts knowledge for policy-making, consider deforestation in developing countries. A common scenario is that forest-owning households cutting down their trees do not take account of the external cost of their contribution to global warming. A solution has long been known, namely a Pigouvian tax. But this would be hard to implement as an RCT, since the power to tax mostly lies with governments, who would (understandably) be resistant. Instead, one can randomize payments to those who choose not to cut down their trees, as in the RCT for Uganda by Jayachandran et al. (2017). This policy can be implemented by a local NGO, bypassing the government. Here there is a public-economic rationale, but the use of an RCT constrains the policy options evaluated. And the tax policy will probably have different impacts (if only because the payment policy gives extra value to the stock of trees, generating an income effect, separately to the price effect).

Of course, no single tool can cover all applications. The question here is whether we have a reasonable balance today between research effort and policy challenges. The (questionable) hierarchy of methods advocated by the randomistas makes it harder to attain that balance. Indeed, even for private goods, the very idea of randomized assignment is antithetical to the goals of many development programs, which typically aim to reach certain types of people or places. In delivering cash

transfers to poor people—a favorite intervention for development RCTs—governments will hopefully be able to do better than a random assignment.

The aforementioned IEG report documents the unbalanced assignment of World Bank impact evaluations across the sectors of its operations, and the seemingly poor fit of the evaluation portfolio to the Bank's sectoral and development priorities (World Bank 2012). Though I have not seen evidence, I suspect that there is also an imbalance in the assignment of evaluative effort according to the likely duration of project benefits. Long-term evaluations of World Bank development projects are rare, despite the claims made about longer-term impacts. I can testify from personal experience how hard it is to organize and implement long-term evaluations at the World Bank.[66] It is plausible that favoring RCTs exacerbates a myopia bias in development knowledge.

This is not just happening in the World Bank. The sectoral bias in the use of RCTs more broadly is evident from the results of Cameron et al. (2016) who provide a cross-tab of over 2200 published impact evaluations (in the aforementioned 3ie database) by method and sector.[67] Overall, about two-thirds of these evaluations use RCTs, but the RCTs tend to be concentrated in certain sectors, notably education (58 percent used an RCT), health, nutrition, and population (83 percent; 93 percent in health alone), information and communications technology (67 percent), and water and sanitation (72 percent). Observational studies are more common—with under one-third using an RCT—in agriculture and rural development, economic policy, energy, environment and disaster management, private sector development, transportation, and urban development. The production of impact evaluations has also been uneven geographically (even allowing for population). India has had the largest absolute number but Kenya has had the most per capita.[68] The geography of RCT placement is influenced by researcher connections with local NGOs.

There are both supply and demand sides to this bias. On the supply side of evaluations, the reality today is that, enamored by the promise of cleanly identifying a causal effect, many economists and other social and political scientists have been searching for something to randomize. If randomization is not feasible, they turn to ask another question.

On the demand side, governments (and development agencies) are largely free to choose what is evaluated. One concern here is that they do not always know what evidence they need (Duflo 2017). Politics also plays a role. They may be drawn to pick programs for which there is little risk that a negative appraisal will

[66] This largely based on the study reported in Chen, Mu, and Ravallion (2009).

[67] In addition to RCTs the methods identified are difference-in-differences, instrumental variables, regression discontinuity and matching. Multiple methods are allowed in the counts.

[68] For details see Cameron et al. (2016) and Sabet and Brown (2018).

hurt politically, or to pick those that do matter but for which there are good reasons to be confident of a politically acceptable result (again raising ethical concerns). Other important programs will not be evaluated. The risks are plain.

Addressing these concerns calls for more strategic evaluation agendas, not driven by the methodological preferences of researchers. We have started to see more strategic agendas for RCTs. This is welcome, though the strategies are still led by academic researchers, based on their interests and devoted to one tool. If we are really concerned about obtaining unbiased estimates of the impact of the portfolio of development policies it would surely be better to carefully choose (or maybe even randomly choose!) what gets evaluated, and then find the best method for the selected programs, with an RCT as only one option. That is what is called for if we take seriously the goal of obtaining an unbiased assessment of overall development impact. Research can serve that goal, but it is unlikely to happen automatically.

1.6 Conclusions

We are seeing a welcome shift toward a culture of experimentation in fighting poverty, and addressing other development challenges. RCTs have a place on the menu of tools for this purpose. However, they do not deserve the special status that advocates have given them, and which has so influenced researchers, development agencies, donors, and the development community as a whole. To justify a confident ranking of two evaluation designs, we need to know a lot more than the fact that only one of them uses randomization.

The popularity of RCTs has rested on a claimed hierarchy of methods, with RCTs at the top, as the "gold standard." This hierarchy does not survive close scrutiny. Despite frequent claims to the contrary, an RCT does *not* equate counterfactual outcomes between treated and control units. The absence of systematic bias does not imply that the experimental error in a one-off RCT is less than the error in some alternative non-random method. We cannot know that. Among the feasible methods in any application (with a given budget for evaluation), the RCT option need not come closer to the truth. Indeed, if the sample size for an observational study is sufficiently greater than for an RCT in the same setting, then the trials by observational study can be more often close to the truth even if they are biased.

There is still ample scope for useful observational and other non-random studies (such as deterministic experimental assignments), informed by theory. Yes, there is model uncertainty, though generally not as much as the randomistas assume. Moreover, when we look at RCTs in practice, we see them confronting problems of mis-measurement, selective compliance and contamination.

Then it becomes clear that the tool cannot address the questions we ask about poverty, and policies for fighting it, without making the same type of assumptions found in observational studies—assumptions that the randomistas promised to avoid.

RCTs are also ethically contestable in a way that observational studies are not. The ethical case against RCTs cannot be judged properly without assessing the expected benefits from new knowledge, given what is already known. Review boards need to give more attention to the ex-ante case for deliberately withholding an intervention from those who need it, and deliberately giving it to some who do not, for the purpose of learning. There may be a good case in specific contexts, based on the limitations of existing knowledge, but the case does need to be made in a credible way and not just taken for granted.

The questionable claims made about the superiority of RCTs as the "gold standard" have had a distorting influence on the use of impact evaluations to inform development policy-making. The bias stems from the fact that randomization is only feasible for a non-random subset of policies. When a program is community- or economy-wide or there are pervasive spillover effects from those treated to those not, an RCT will be of little help, and may well be deceptive. The tool is only well suited to a rather narrow range of development policies, and even then it will not address many of the questions that policy-makers ask. Advocating RCTs as the best, or even only, scientific method for impact evaluation risks distorting our knowledge base for fighting poverty. That risk was one of the main concerns in Ravallion (2009a), and the experience since then has reinforced that concern.

While we have seen much progress over the last ten years, there are still grounds for doubting whether evaluative research on development fits well with the policy challenges now faced. This chapter has argued that a better alignment requires:

- Abandoning claims about an unconditional hierarchy of methods, with RCTs at the top, and making clear that "scientific" and "rigorous" evidence is not confined to RCTs.
- Demanding a clear and well-researched ex ante statement of the expected benefits from an RCT, to be weighed against the troubling ethics.
- Making explicit the behavioral assumptions underlying randomized evaluations, similarly to the standards of structural approaches.
- Going beyond mean causal impacts, to include other parameters of policy interest and better understanding the mechanisms linking interventions to outcomes.
- Viewing RCTs as only one element of a tool kit for addressing the knowledge gaps relevant to the portfolio of development policies.

Acknowledgement

François Roubaud encouraged the author to write this chapter. The author thanks Jason Abaluck, Sarah Baird, Radu Ban, Mary Ann Bronson, Caitlin Brown, Sylvain Chabé-Ferret, Kevin Donovan, Ryan Edwards, Markus Goldstein, Miguel Hernan, Emmanuel Jimenez, Max Kasy, Madhulika Khanna, Nishtha Kochhar, Agnès Labrousse, Andrew Leigh, David McKenzie, Rachael Meager, Berk Özler, Dina Pomeranz, Lant Pritchett, Milan Thomas, Vinod Thomas, Eva Vivalt, Dominique van de Walle, Andrew Zeitlin, and participants at an authors' workshop in Paris, March 2019. The staff of the International Initiative for Impact Evaluation kindly provided an update to their database on published impact evaluations and helped with the author's questions.

2
Randomizing Development
Method or Madness?

Lant Pritchett

2.1 Introduction

Bill Gates has recently been promoting chicken ownership to address poverty in Africa. In an open letter, Professor Blattman of University of Chicago pointed out that cash transfers may be more cost effective than chickens: "It would be straightforward to run a study with a few thousand people in six countries, and eight or 12 variations, to understand which combination works best, where, and with whom. *To me that answer is the best investment we could make to fight world poverty.* The scholars at Innovations for Poverty Action who ran the livestock trial in Science agree with me. In fact, we've been trying, together, to get just such a comparative study started."[1] [emphasis added]

I think it is important for the development community to stop and reflect on how we, as a development community, arrived at this two-fold madness. First the madness that Bill Gates, a genius, a humanitarian, an important public intellectual, could be even semi-seriously talking about chickens. Second, the madness about method, that the response of Chris Blattman, also a genius, an academic at a top global university, and also an important public intellectual, would respond not "Chickens? Really?" but rather that the "best investment" to "fight world poverty" is using the *right method* to study the competing program and design elements of chickens versus cash transfers.[2]

That this *is* madness is, I hope, obvious. The top 20 most populous developing countries in the world are (in order): China, India, Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, Russia, Mexico, Philippines, Ethiopia, Vietnam, Egypt, Iran, Turkey, DR Congo, Thailand, South Africa, Tanzania and Colombia. Together these countries have 4.6 billion people. Imagine gathering a couple of dozen of

[1] https://www.cgdev.org/blog/getting-kinky-chickens

[2] With dozens on studies on conditional cash transfers, microfinance, and a sobriquet "Worm Wars" to describe a massive debate on whether deworming is cost-effective (and a bouquet of RCT studies of boutique anti-poverty and kinky goal interventions) this madness has seeped far more broadly.

Lant Pritchett, *Randomizing Development: Method or Madness?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud,
Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0004

the leaders from any one of these countries (where "leadership" could be political, social, economic, intellectual, popular, mass movement, civil society, or any combination) and saying: "We, the experts in the development community, think 'fighting world poverty' is the center of the development agenda and we think that the 'best investment' we can make to promote development/fight poverty in *your country* [fill in the blank: Indonesia, Brazil, Nigeria, DRC, Tanzania, South Africa, Egypt, India] is a set of studies using the right method to resolve the questions of whether anti-poverty programs should promote chicken ownership or distribute cash and, within that, how best to design such chicken or cash transfer programs?"

I imagine two responses from country leaders. One, how could you have come to such trivial and trivializing ideas about our country's goals, aspirations, and challenge? How can we as [Indonesians/Indians/Nigerians/Egyptians/Tanzanians] not take as outright contempt the suggestion that either "chickens" or "studies about chickens" are the top priorities for our country? Two, we can easily list for you many pressing, urgent, if not crisis, development issues affecting the current and future well-being of the citizens of our country. These questions are important whether or not your preferred method for producing research papers can address them.[3]

I am using "studies of chickens versus cash" not to single out Professor Blattman, but to stand in for the whole *randomista* movement in development. Development economists, rather than finding it hard to think of "anything else" (Lucas 1988) but the big picture issues around national development, are now so committed to a method they are thinking about "anything but" national development. There are now literally thousands of published RCTs, with dozens on studies on conditional cash transfers, on microfinance, and literally hundreds of studies of boutique

[3] Four (of many possible) anecdotes to back this assertion up. First, a colleague of mine was in the front office of the prime minister of a large and important country. At the request of prominent randomistas who had done considerable work in that country he managed to set aside two hours for a meeting between these academics and the prime minister. At the end of the meeting the prime minister pulled my friend aside and said: "Never, ever, waste my time like that again." Second, my colleague Arvind Subramanian was a top policy adviser in India, a country that has been a focus of randomistas activity, for three years. In a speech to my students in 2018 he said that *never* in his three years of being involved at many levels (from mid-level to the highest) in discussing the range of economic challenges facing India did he hear the results of any RCT play any role. Third, in my work as a development practitioner I have been in all but two of those top twenty population countries and have lived for years in two of them (Indonesia and India) and never, ever, outside of the narrow confines of development agencies and projects have I heard either chickens or rigorous studies mentioned as priorities. Fourth, when the "livestock trial in Science" study was being promoted in the media a reporter from a US-based publication called to ask me my view of this important study. I responded that I had not read it as it wasn't a particularly interesting or important study from my viewpoint as a development scholar/practitioner. She asked me how, in light of the august authors and preeminent publication I could say such a thing. I responded that if she could find any mention of that study in the local press or media in *any* of the seven countries I would change my mind, read the study, and give her comments. Since of course the reporter never called back, I had a research assistant search for media mentions in any of the study countries (canvassing for people who spoke the local languages to help) and we could not come up with a single local media mention of the study.

interventions in water, sanitation, education, health, business training, etc.[4] I argue this madness about a method in development academia is a symptom, not the disease. The big debate is about the relative importance of "national development" versus "kinky development" and whether "national development" can be accelerated. RCT as a method can only even pretend to any importance if either (a) one interprets the development in a narrow way as achieving specific, low-bar, targets ("kinky development"), or (b) one takes the view that "national development" is completely beyond the influence of ideas or evidence.

National development is a four-fold transformation of an *intrinsically social* grouping (country or region or society) to higher levels of capabilities in four dimensions: an economic transformation from lower productivity to higher productivity; a political transformation to governments more responsive to the broad wishes of the population, an administrative transformation to organizations (including those of the state) with higher levels of functional capability for implementation, and a social transformation to more equal treatment of the citizens of the country (usually with a sense of common identity and, to some extent, shared purpose). National development is about countries like Haiti or India or Bolivia or Indonesia achieving the high levels of economic, political, administrative, and social *functional* capabilities that Denmark or Japan or Australia possess. National development is not an end but a means of achieving a higher level of human well-being.

> "Kinky development" (Pritchett 2014a, Kenny and Pritchett 2013) is the view that development is primarily, if not exclusively, about reaching very low-bar levels of specific indicators: "eradicating extreme poverty" or "universal primary school completion" or "access to safe water" are "kinky" goals in that they draw some completely arbitrary line or threshold in some dimension of human well-being and then pretend that "kinking" the distribution of well-being, pushing people to just that threshold, is the goal of development. The distinctive element of kinky development is that gains to human well-being above the low-bar threshold count for *nothing*.

Section 2.1 empirically demonstrates two things.

One, median income/consumption, one of the four elements of national development, is both (a) empirically *necessary* and *sufficient* for reducing headcount consumption poverty and (b) accounts for that *essentially all* of the cross-national variation in poverty rates. The effect of anti-poverty programs (and a fortiori the design of such programs and a fortiori squared, so to speak, studies

[4] There is even a term "Worm Wars" to describe a hotly contested debate on the questions of whether, when and where, deworming is a cost-effective intervention.

about the design of anti-poverty programs) are just tiny compared to the effects of inclusive growth.

Two, for omnibus measures of human well-being, such as the Social Progress Index, (a) high levels of national development are empirically *necessary* and *sufficient* for achieving high levels of human well-being and (b) this relationship is empirically tight for the Social Progress Index (and other omnibus human well-being measures). Moreover, all (less one) of the dozen of specific measures of human well-being that go into the Social Progress Index (e.g. access to water, personal security, health, education, etc.) are also tightly correlated with national development.

Section 2.2 presents a decision-tree framework to evaluate the claim that a specific intellectual activity (such as an RCT study) about targeted programs (like cash versus chickens) could be the "best investment" for "fighting poverty" (or, more generally, any measure of human well-being). I show that *all* the links in the chain of reasoning that are needed to arrive at such a conclusion are false.

2.1 National Development and Human Well-Being

I propose a rough and ready definition and empirical measures of "national development" and then show its empirical relationship to measures of human well-being, both kinky measures, like low-bar poverty, and broader measures.

2.1.1 National Development as a Four-fold Transformation of Countries

The very word "development" implies a change over time in which something becomes a better, more mature, more advanced version of its ontological type. A human develops from zygote to mature adult, a frog from zygote to tadpole to frog. Rocks neither "develop" to become frogs nor do rocks, through erosion, "develop" to become sand. The first is impossible and the latter not directional. What is it that "develops" with "development"? With "national" development what "develops" is typically a country, but is always and intrinsically a *social* (and socially constructed) aggregate.[5] A country has (at least) four important dimensions along which it "develops" and each is *intrinsically* and *ontologically* social and cannot be meaningfully individuated.

[5] While "nation" or "nation-state" are often used casually as synonyms for "country" this language brings in massive ideological baggage about what a "nation" is and its relationship to sovereign states as "countries." We can talk about the "development" of regions (e.g. Southern versus Northern Italy) or of provinces/states within a country (e.g. Tamil Nadu versus Uttar Pradesh).

Economic development. This is usually understood as the productive capability of a *place*. This has some elements of the characteristics of the individuals but also a general "total factor productivity"-like element which is place-specific and not individuated. A country's labor productivity, as measured by GDP per worker, is one possible indicator of economic development, though there can be many others (e.g. Hidalgo and Hausmann (2009) measures of economic complexity), and GDP can be adjusted in many ways (e.g. green accounting). These measures are *never* intended as direct measures of human well-being but are measures of the economic product and productivity of a place.

Administrative development. This is typically conceived of as some aggregate of the capability of (mostly state) organizations to accomplish public purposes.[6] Countries have an array of organizations to carry achieve purposes: armies, central banks, post offices, police forces, courts, land registries, etc. While there is of course variation within countries in the capability of organizations (Kaufmann, Mehrez, and Tugrul 2002), an aggregate of the administrative capability of the state is another element of national development. The Fragile Sates Index, as one example of such a measure, ranks countries from 0 (best) to 10 (worst, most fragile) on their "broad based provision of public services" and Denmark scores 0.9, Indonesia 5.6, and Haiti 9.4.

Political development. This is obviously hugely value laden and, like anything said about politics, is itself political, but *descriptively* when people described the "development" of states they usually had in mind some notion that those in political power and exercising sovereign power in a country: (a) are responsive to the needs, wishes, wants, desires of the citizens of the country and that political processes allowed those to be expressed by citizens and aggregated in fair and legitimate ways and (b) respected at least some set of "negative" rights that preserved liberty and security of the person (and perhaps in addition some "positive" rights) and (c) there is some degree of "rule of law." The Fragile States Index, for instance, has two distinct measures, one for "state legitimacy" (not "democracy") and one for "human rights and rule of law" (10 is worst, 0 is best). For State Legitimacy Haiti is 8.7, Indonesia 4.8, and Denmark 0.9 while for Human Rights and Rule of Law Haiti is 7.4, Indonesia is 7.3, and Denmark is 1.2. The Polity2 measure of the POLITY IV project is on a plus 10 to minus 10 scale where 10 is complete democracy and minus 10 is complete autocracy. For example, the measure has been 10 for Denmark since 1915 (with the interregnum of WWII); in Indonesia was -5 in

[6] In our work *Building State Capability* we distinguish between the *capability* of organizations, which is a feature of an organization, and *capacity* as a feature of individuals and point out that capability or an organization is not the aggregation of the capacity of the individuals. This is to emphasize there are two distinct concepts, but we acknowledge one could just as well have used the words exchanged (e.g. capacity as a feature of organizations) and, as long as one were consistent about distinguishing the two concepts, achieve the same goal.

1998 (last year of Suharto's rule), jumped to 6 in 1999, and was 9 by 2017; in Haiti this was 0 from 2010 to 2015 and 5 in 2016 and 2017.

Social development. Even more value laden and hence, if anything, more political than political development, is the notion of how citizens/members of a common society treat each other changes as an intrinsic part of development. While these ideas were flawed in many ways (and in many ways reprehensible projections of social constructs of colonialists and colonialism) there was an important notion that "social equality"—in the sense that people were treated by other people equally independent of their social identities (kin, hereditary class, clan, tribe, ethnicity, race, sex, religion)—was, in and of itself, part of development. One part of the social development was the creation/adoption of a shared identity. These are obviously historically constructed values of the Western experience and do not have universal validity, but I would argue were often bundled into notions of "modernization" and "development" for good or ill. Today of course this is most obvious in the views that development needs to be gendered and that societies that do not treat the sexes fairly are considered less "socially developed" at least in one important sense, than those that do.

The units at which national development happens: a market, an organization, a polity, a society are about processes in which individuals participate and into which they are embedded but are *ontologically* not individuated.

2.1.2 Levels of Median Income/Consumption Completely Explain Poverty

National development, and in this case, just one measure of one element of national development, the levels of median consumption, is *sufficient* to (essentially) eliminate "low bar" or "dollar a day" (now, with inflation, P$1.90 a day where "P$" means purchasing power adjusted dollars) poverty. The standard World Bank data, limited to all country/year pairs with actual survey data, one has over 800 country/year observations on measured poverty rates and on median income or consumption. Figure 2.1 shows that *no country* with median annual income above P$3,000 (about the level of Peru or Mongolia around 2010) has low-bar poverty more than 10 percent. By P$5,000 (about the level of Costa Rica) essentially no country has low-bar poverty above 2 percent. Also, no country with median income above P$1,000 (about the level of Bangladesh in 2010) has low-bar poverty more than a third of their population. The white-space in the "northeast" of Figure 2.1 is important as those are combinations of median income/poverty that *never* happen. There is a level of median income/consumption that is empirically *sufficient* to reduce poverty below any given percent of the population.

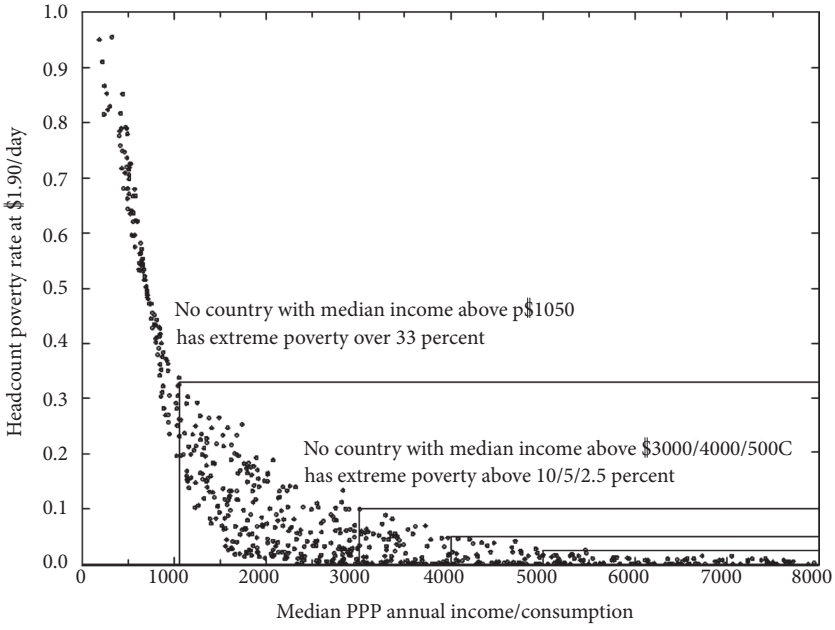

Figure 2.1 Median income/consumption is sufficient to eliminate extreme poverty

Source: Author's calculations with data from PovcalNet: the online tool for poverty measurement developed by the Development Research Group of the World Bank (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

Figure 2.2 shows the levels of median income/consumption that are *empirically necessary* to reach various levels of $5.5 per day poverty rates.[7] By "empirically necessary," I am not asserting any logical necessity (like a theorem) but just that is doesn't happen. The whitespace in the "southwest" of Figure 2.2 corresponds to low median/low poverty headcount poverty rate combinations that are never seen. No country has pushed $5.5/day poverty below 75 percent of all households without median income above P$1045. That implies 42 of the 164 countries have a latest observed level of income such that *no country* has *ever* been observed with a poverty rate at P$5.5 less than 75 percent with their level of income. 107 of the 164 countries have a level of income such that (almost) no country has been observed with poverty below 10 percent at their level of income. No country (but one[8]) has pushed P$5.5/day poverty below 10 percent without having median

[7] This is the highest level the World Bank source provides data but this is a "moderate" not a "high" poverty line. I, and many other people, argue for upper bar poverty definitions of P$10/day or above, which are still far below those actually used in richer countries.

[8] This country/year is Azerbaijan in 2005, whose data show median income of P$5655 in 1995 and poverty headcount 5.5$/day poverty of 5 percent and median income of P$5197 in 2015 and poverty of essentially zero but in 2005 a median income of P$2785 and poverty of 7.7 percent, which is the anomalously low observation, even for this country.

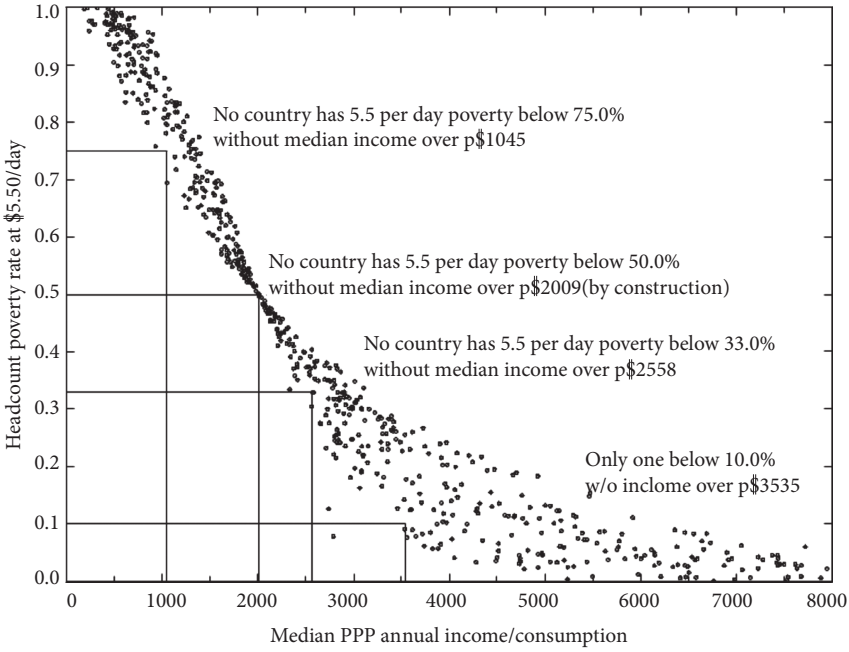

Figure 2.2 High levels of median income/consumption are empirically necessary to eliminate poverty (and these levels are higher the higher the poverty line)

Source: Author's calculations with data from PovcalNet: the on-line tool for poverty measurement developed by the Development Research Group of the World Bank (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

income/consumption above P$3535 (roughly the level of "upper middle income" countries like Peru (P$3486 in 2015), Kazakhstan (P$3557 in 2015), or Thailand (P$3549 in 2010).

So far, I have been using 810 observations from the World Bank data whether the data was for income or consumption. But for exploring connections with programs or projects consumption expenditures are a better measure as they more reliably measure post-tax and transfer outcomes and hence reflect consumption expenditures inclusive of any benefits from programs. Figure 2.3 shows the relationship between country level poverty rates at the three poverty lines in the World Bank data, P$1.9, P$3.2, and P$5.5, and the median of the distribution of consumption using just the 389 country/year observations using consumption data. Since the poverty rates *must be, by construction*, non-linear in the median, I fit a completely flexible functional form including all powers of the median from –2 to 5.

For all three measures the data say that *very nearly all* the observed variation (R2 of 0.983 to 0.988) across countries and time in poverty rates is associated with variation in the median (50th percentile) of consumption. An R2 of 0.988 implies that the correlation of actual poverty rates and the poverty rate predicted from the median is 0.994 (=√.988).

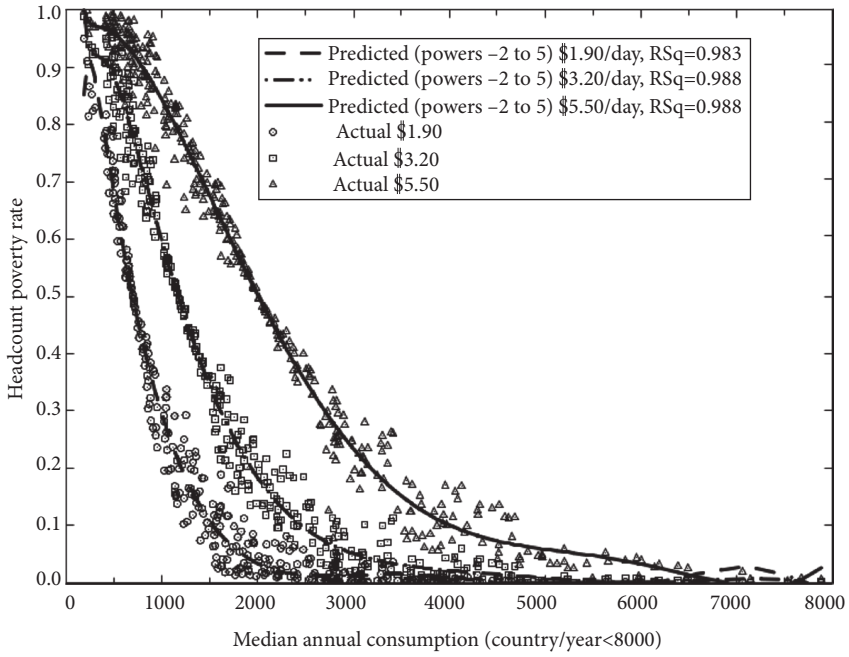

Figure 2.3 Median income/consumption of a country predicts the level of poverty exactly for high poverty lines and near exactly even for low poverty lines

Source: Author's calculations with data from PovcalNet: the online tool for poverty measurement developed by the Development Research Group of the World Bank (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

This of course doesn't mean that other factors like the change in the inequality or the adoption of "poverty" programs *cannot* make a difference or even that they *cannot* in principal make a "substantial" difference, it just says that empirically, relative to the massive changes associated with the change in the median (from poverty of 100 percent to near zero percent), the differences at a given level of consumption are very modest compared to the gains from growth. Table 2.1 shows calculations of various poverty counter-factuals. For a country in the middle of the bottom quartile the poverty rate is 72.2 percent. If the country moved "due south"—had a lower poverty for the same median consumption—by one standard deviation of the residual the poverty rate would be 68.6 percent. In contrast if that country had the median consumption of having grown by 2 ppa faster over the previous 20 years (roughly a standard deviation of cross-national growth rates) its poverty would have been more than halved, to 35.9 percent. It would take a growth rate only 0.2 percent higher (e.g. 2.2 ppa vs 2 ppa)—which is only a tenth of a cross-national standard deviation—to produce the same poverty reduction as improving poverty for a given median by a standard deviation.

Table 2.1 Even very small improvements in growth produce poverty reduction near the same as substantial (standard deviation of residual) improvements in poverty for a given level of median consumption

Poverty rate	Quartile I of consumption, $1.90/day poverty line	Quartile II, $5.50/day
At average median consumption in the country quartile	72.2%	74.1%
If poverty is one standard deviation of the residual better for same consumption	68.6%	70.2%
If medium run growth (20 years) were 2.0 ppa higher (one cross-national standard deviation of growth rates)	35.9%	51.8%
If medium run growth (20 years) is better by 0.2 ppa (one tenth of a cross-national standard deviation of growth rates)	67.8%	72.2%

Source: Author's calculations with regressions shown in Figure 2.3.

This super-tight correlation of measured poverty rates and median income/consumption also hold in changes over time within countries (Kraay 2006).[9] Figure 2.4 shows an R2 of 0.93 between the change in "dollar a day" (P$1.90) poverty with the change in the predicted poverty based on just the shift in the median and the estimated functional form for the longest observed spell (longer than 10 years) for each country.

Figure 2.5 shows some large countries that have seen extreme poverty fall rapidly from very high levels to low levels: China, Indonesia, Vietnam and, to a lesser extent, India. These poverty reductions happened right in front of our eyes as we have reasonably good household surveys tracking poverty over most (or all) of these periods and so careful empirical work can be done to decompose the proximate determinants of this fall. How much of this fall in poverty was "accounted for" by changes in the central tendency (mean/median), how much was general change in inequality and how much was due to shifts in the distribution below the poverty line, conditional on mean and overall inequality of the type that "anti-poverty programs could in principle be responsible for). It is not too terrible a caricature of these results to say that "all" or "more than all" of the reduction in poverty in these countries was due to shifts in the mean/median. "More than all" is possible in that in many cases inequality got worse (in the case of China much worse) and hence the increase in the central tendency had to offset that poverty worsening increase in inequality to reduce poverty.

[9] All of the empirical work here relies on the standard World Bank sources on household incomes/consumption, not on estimates of GDP per capita. Pinkovskiy and Sala-i-Martin (2016) argue, based on satellite data of light at night, that GDP per capita is a better, more reliable measure of progress and this shows faster progress and more poverty reduction.

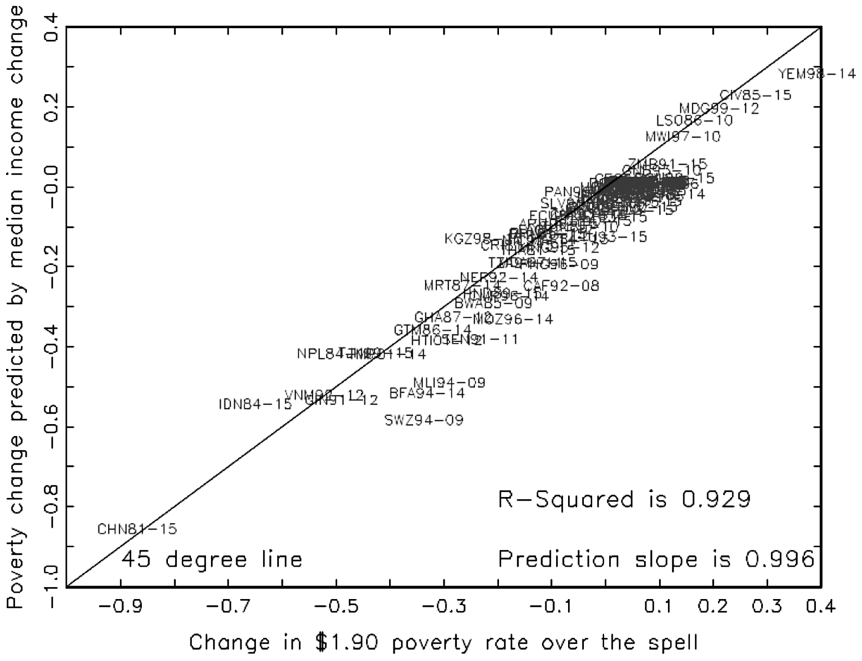

Figure 2.4 Changes in poverty rates are also tightly associated with changes in median income/consumption

Source: Author's calculations with data from PovcalNet: the online tool for poverty measurement developed by the Development Research Group of the World Bank (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

By "poverty programs" many people seem to mean "interventions" that raise the consumption of "the poor" at a given level of the median. This is what would be expected from a cash transfer (conditional or not), a "graduation" type livestock program (referred to above), microfinance, chickens, business training or pretty much any other *targeted* anti-poverty program. These are all intended to bringing up the "left tail" of the consumption distribution (benefitting "the poor") while holding its central tendency fixed (or possibly lowered, depending on how programs are financed). The simple correlations say that differences across the country/years in the impact of "poverty programs" conditional on the median account for *at the very most* 1.2 percent of the total cross-national variation in poverty rates.[10] This is an upper bound as *everything* besides the median (measurement error, non-programmatic differences in left-tail versus median

[10] The standard poverty measures and medians are just different summary statistics of the *same* distribution. The standard headcount measure is just a partial integral of the distribution below a poverty line (I have published papers on methods for calculating poverty, e.g. Pradhan et al. 2001). This doesn't mean a high correlation is "baked in" as it would be possible, in theory, for programs to "kink" the distribution and reduce poverty for a given median.

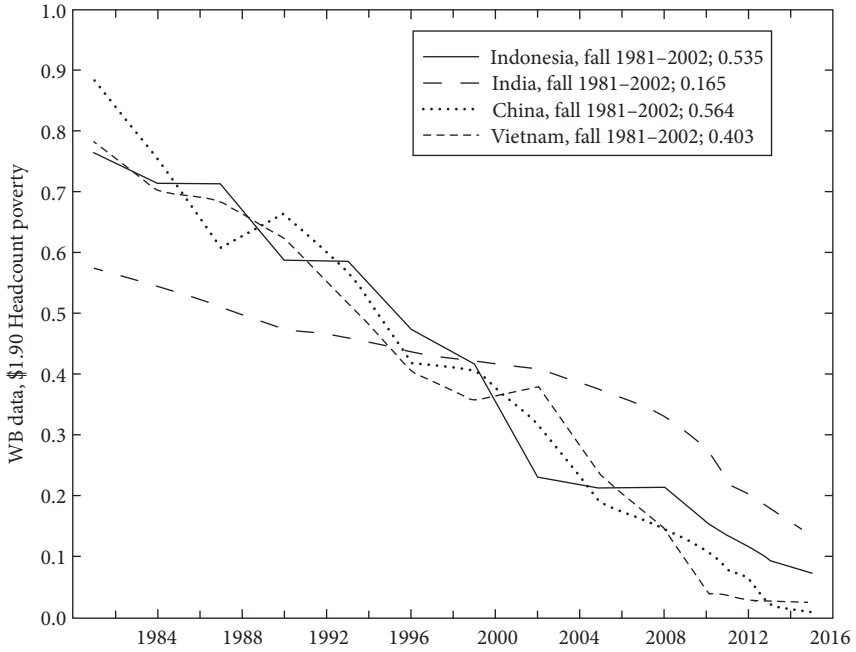

Figure 2.5 In several countries the most rapid reductions in extreme poverty in history had been underway for 20 years by 2000

Source: Author's calculations with data from PovcalNet: the online tool for poverty measurement developed by the Development Research Group of the World Bank (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

consumption (e.g. different relative prices of goods the poor consume intensively), non-programmatic differences in incomes driven by different relative prices of assets owned by the poor (e.g. unskilled labor), etc.) adds up to 1.2 percent of the observed variance in poverty so poverty programs could account for as little as 0.1 percent (given the existence of scaled and effective programs in at least some places, it is unlikely to be exactly zero).

2.1.3 National Development and Broader Measures of Social Progress

In addition to its impact on a kinky goal like extreme poverty, achieving high levels of national development is also a necessary and sufficient condition for achieving high levels of overall human well-being. The correlation of an omnibus measure of human well-being (Social Protection Index) and national development are extremely high (.967) (Pritchett 2016).

The Social Progress Index[11] is the result of the effort of the Social Progress Imperative to create a new and better ways to compare development performance across countries. They explicitly do *not* use GDP per capita (or other measures of national development), but rather focus on direct measures of human well-being. The Social Progress Index (SPI) has three aggregate components called: (1) basic human needs, (2) foundations of well-being, and (3) opportunity. Each of these three components are built from four subindicators, which are each themselves built up from specific measures. For instance, the aggregate "basic human needs" (I) has four subcomponents: I.1 "nutrition and basic medical care," I.2 "water and sanitation," I.3 "shelter," I.4 personal safety. Each of these is based on specific indicators, so, for instance, subcomponent I.2 "water and sanitation" is based on: I.2.a "access to piped water," I.2.b "rural access to improved water source," and I.2.c "access to improved sanitation." I am not saying the SPI is the best measure of country-level human well-being, but it is a thoughtful and careful attempt to measure social progress across countries and uses 53 distinct indicators—which include economic, education, and health indicators but also non-standard indicators like religious tolerance, freedom from crime, and political rights.

I regress the SPI (re-scaled 0 (worst) to 100 (best)) on three indicators of national development: (ln) GDP per capita (proxy for productive economy), the POLITY2 measure of autocracy/democracy (proxy for responsive polity), and World Governance Indicator of Government Effectiveness (proxy for capable administration), also each scaled 0 to 100[12] for 140 countries (excluding high income oil countries and one country (El Salvador) whose GDP per capita data seemed wrong). The National Development Index adds the three components using OLS coefficients as weights.

Figure 2.6 shows that national development is empirically necessary and sufficient for achieving high levels of the SPI. No country has achieved an SPI in the top third of countries (above 70.1) without a National Development Index above 68.6 (Argentina's level).[13] Similarly, no country in the top third of NDI (National Development Index) has an SPI less than 61.6.

[11] http://www.socialprogressimperative.org/global-index/

[12] I don't think any hinges on using these particular three proxies for the underlying concepts of national development. For instance, the Fund for Peace presents a Fragile States Index that has multiple components. Two of those, "Public Services" and "State Legitimacy" are potential alternative empirical proxies for the concepts of "administrative capability" and "political responsiveness." A regression of the overall Social Progress Index on GDP per capita, FSI: Public Services and FSI: State Legitimacy (all scaled to 100) the R-Squared is 0.947 (even higher), with all three indicators having powerful roles.

[13] Measures of human well-being are sometimes to point out that GDP per capita is a weak proxy for human well-being (for which, of course, no economist ever proposes it) by showing "outliers" that achieve high SPI with low(ish) GDP per capita. But "national development" *includes* politics, state capability, and social transformation. With this broader definition countries that are sometimes high performers for their GDP per capita like Costa Rica (CRI in the graph, which overlaps URY, i.e. Uruguay) does have high SPI and "over-performs" even its NDI, but it is not a massive "outlier" as it has high NDI.

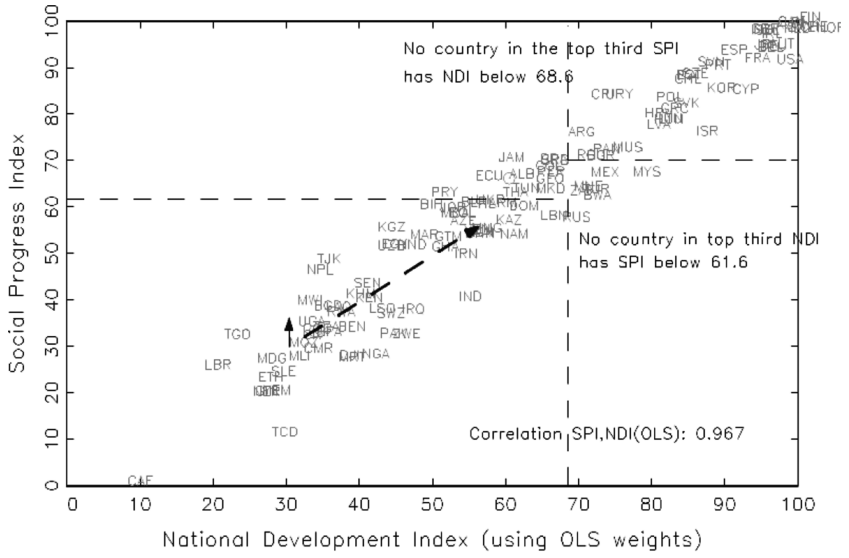

Figure 2.6 National development is empirically necessary and sufficient for high levels of the Social Progress Index

Source: Author's calculations with data and procedures as described in the text.

The SPI and NDI have a correlation of 0.967 (R2 of the regression was 0.935). This is an amazingly tight relationship of two conceptually and empirically different measures as different cross-national measures of the *same thing* from different sources or methods—like "years of schooling of the adult population" or "child mortality"—often don't have cross-national correlations as high as 0.96, just due to pure measurement error.

As with poverty, the strong and tight relationship implies the potential gains in social progress for a given level of national development are quite small (relative to the range of SPI). Mozambique (abbreviation MOZ) has roughly the same actual and predicted SPI (hence NDI) of about 30. Suppose somehow Mozambique were a "star performer" on Social Progress for a given level of national development, in the specific sense it has SPI higher by a residual standard deviation (so, on the assumption of a normal distribution was in the 84th percentile of countries with its NDI rather than 50th). Then its SPI would be 36 (illustrated with the vertical arrow in Figure 2.6). This gain is not nothing, but still would leave Mozambique's SPI below Laos, Bangladesh, or Kenya. In contrast, if Mozambique improved by one standard deviation on each of the elements of national development the SPI would reach 56, higher than the SPI of upper-middle income countries like Morocco or Indonesia (dashed "northeast" arrow in Figure 2.6).

Table 2.2 shows the empirical relationship of the three components and 12 subcomponents of the Social Progress Index with proxies for national development. Each of the three components of the SPI has a very strong correlation with NDI

Table 2.2 The Social Progress Index—and all of its components and subcomponents—are strongly associated with three indicators of national development

Social Progress Indicator, its three components (Basic Human Needs, Foundations of Well Being, Opportunity) and the four subcomponents of each component	Economic Productivity ((ln) GDP per capita, PWT8.0, rescaled 0 to 100)		Administrative Capability (World Governance Indicators, Government Effectiveness, rescaled 0 to 100)		Political Responsiveness (Polity IV Project, Polity 2, rescaled 0 to 100)		R-Squared of regression on national development indicators
	OLS coeff.	t-stat.	OLS coeff.	t-stat	OLS coeff.	t-stat	
Social Progress Index	0.53	13.67	0.34	7.38	0.12	5.01	0.935
I) Basic Human Needs	0.74	12.10	0.18	2.46	−0.02	−0.43	0.835
I.1) Nutrition and Basic Medical Care	0.57	8.86	0.34	5.17	0.18	5.06	0.865
I.2) Water and Sanitation	0.31	4.95	0.51	8.15	0.23	7.11	0.873
I.3) Shelter	0.80	9.74	−0.09	−0.95	0.04	0.79	0.672
I.4) Personal Safety	1.17	11.78	0.01	0.06	0.06	1.12	0.784
II) Foundations of Well-Being	1.06	13.30	0.04	0.47	−0.01	−0.36	0.820
II.1) Access to Basic Knowledge	−0.02	−0.27	0.77	7.86	−0.09	−1.83	0.603
II.2) Access to Info and Comm.	1.00	10.62	−0.11	−1.09	0.04	0.73	0.707
II.3) Health and Wellness	0.53	8.02	0.22	3.25	0.21	6.11	0.816
II.4) Environmental Quality	−0.18	−1.55	0.50	4.34	0.01	0.13	0.242
III) Opportunity	0.11	1.33	0.52	6.43	0.18	4.34	0.709
III.1) Personal Rights	−0.08	−0.86	0.53	5.68	0.55	11.58	0.765
III.2) Personal Freedom and Choice	0.16	2.06	0.66	8.65	−0.01	−0.37	0.757
III.3) Tolerance and Inclusion	0.19	1.71	0.41	3.70	0.14	2.48	0.517
III.4) Access to Advanced Education	0.93	11.21	0.17	2.04	0.03	0.73	0.824

Source: Author's calculations.

(Basic Needs 0.904, Foundations of Well-Being 0.925, and Opportunity 0.932). All of the 12 subcomponents (less one[14]) are also strongly associated with national development.

National indicators of subjectively assessed well-being are also highly correlated with national development. Regressing the Cantril "ladder of life" measure of average subjective well-being on the three national development indicators has an R2 of 0.66 (correlation 0.812 with an OLS NDI). The World Happiness Report has developed another index of human well-being based on the empirical relationship of seven factors (like "perceptions of corruption," "healthy life expectancy," "social support," and measures of affect) to the "ladder of life" measure of subjective well-being. An equally weighted index of the six elements of the happiness index regressed on the three indicators of national development produces an R2 across 120 countries of 0.788 (correlation with OLS NDI 0.887). Again, the correlation between this six element "happiness" index and the directly observed "ladder of life satisfaction" measure is 0.81. While these are lower than the SPI/NDI correlation, the three indicators of human well-being (SPI, subjective life satisfaction, World Happiness Report) are only about as tightly correlated among themselves as each is with a (measure specific) national development index.

2.1.4 National Development Brings Elimination of Poverty and High Levels of Human Well-being

With the accumulation of more and better data, we can show that the relationships of national development with poverty, overall human well-being, or specific indicators of well-being are as high and tight as anyone ever claimed they would be.

What is odd is that anyone ever doubted this. Four-fold national development is a human well-being machine. Take any objective that contributes to well-being that is strong and widely spread—access to water, better health, improved shelter, more schooling—national development is built to increase the accomplishment of that objective. A more productive economy that produces broad based increases in incomes allows households more income to pursue their objectives so, to the extent these objectives are private goods, it would be very strange indeed if higher private incomes did not lead to higher levels of consumption (and indeed all that empirically matters in the SPI components for "water and sanitation" and "shelter" and "access to basic knowledge" as the only significant correlate is GDP per capita).[15]

[14] The indicator without a strong positive correlation is "environmental quality," which includes greenhouse gas emissions, which are positively associated with GDP per capita.

[15] And, one would expect the relationship with national development to be even stronger/tighter

But if the human well-being objectives require "public goods" (non-rival, non-excludable) or the markets for these goods have "market failures" then this is precisely what governments that are responsive and capable can address. Indeed, for the component "environmental quality" the only strong partial correlates were capability and polity, not GDP per capita and for "personal safety" the only partial correlate was state capability. No one, even the most ardent and market-oriented economist, ever made the case income alone would solve all problems. A responsive polity and capable state was always an integral part of the vision of development.

2.2 RCTs in Development as a Method for Improving Human Well-being

Back to the madness. How did we get to studies of chickens? How did development economics get to thinking about anything *but* national development? How would one provide argumentation or evidence or warrant for a claim that a study with a particular method of the relative effectiveness of targeted programs of chickens versus cash was the "best investment" for fighting poverty? There are three multiplicative elements to such a claim: (a) the likelihood a study produces reliable and useable knowledge, (b) the likelihood the knowledge changes events in the world that improves outcomes, and (c) the total gain to human well-being (in some normative evaluation) from such changes (see Figure 2.7).

Figure 2.8 from the top down provides a map of the array of framings of measures of human well-being (omnibus/aggregate and specific indicators or domains) and whether the normative evaluation of those is kinky or not. The essence of the "kinky" measure is *not* that the poorer (those with less sanitation/education/energy) receive more weight in the measure of human well-being and the richer (those with more of a specific thing) less weight. Any of the standard inequality measures of aggregates, like the Atkinson index or a standard Social Welfare Function with the assumption of declining marginal utility can accommodate that (with parameters giving different intensities of "preference for the poor") and similarly sector measures can give greater weight to specific levels of service or certain groups. The essence of a kinky measure is that the gain to human well-being above some arbitrary threshold (like a poverty line, or "primary school completion" or "access to a latrine") is *exactly zero.*

for "necessities" as economists' definition of "necessity" is something for which marginal utility gets very high as consumption of it falls and, related, something for which the price elasticity (especially at low levels) is expected to be very low. A simple Engel curve—that food share in consumption declines linearly with (log) aggregate income/consumption is arguably the best documented fact in all of economics.

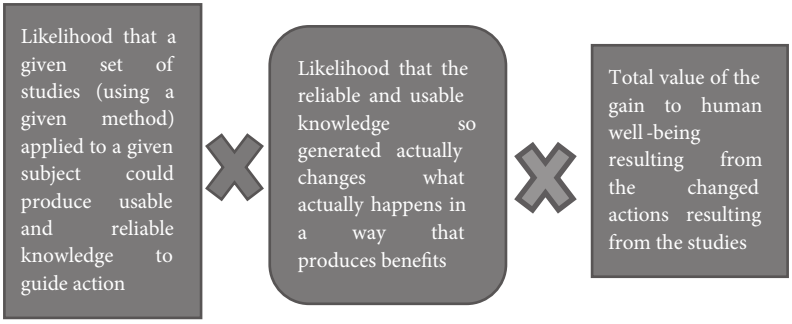

Figure 2.7 The empirical magnitudes to be resolved to make decisions about the expected relative value of various types of investment in research
Source: Author.

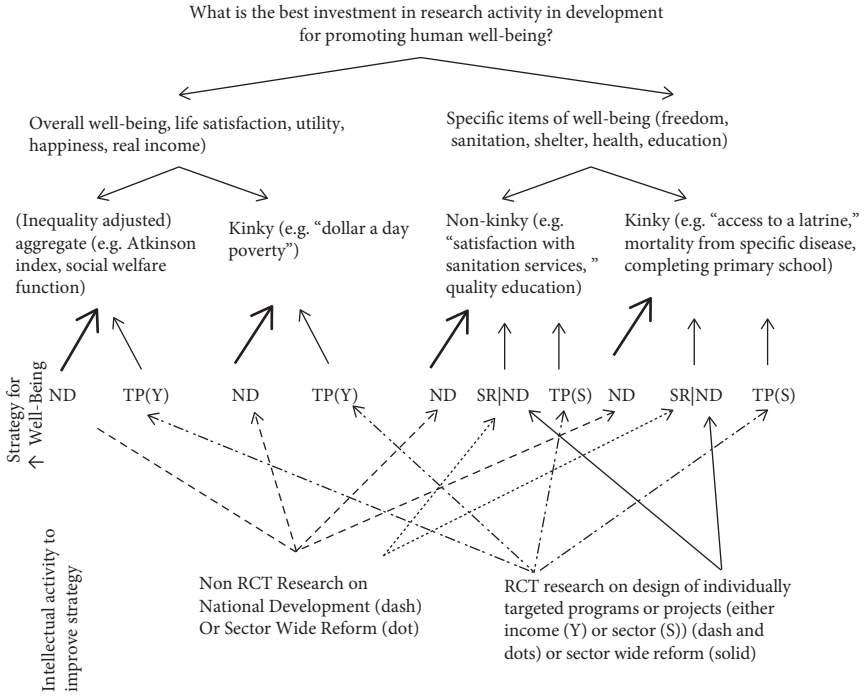

Figure 2.8 What is the best investment in research activity in development for promoting human well-being?
Source: Author.

From the bottom up the arrows illustrate claims about the strength/magnitude of causal impacts on human well-being of national development (ND), targeted programs (in income (TP(Y)) or specific sector indicators (TP(S), or sector-wide reforms (SR|ND).

The "best investment" claim is that the link from RCT study to improved targeted program to raise incomes (TP(Y) (dash and dots arrow) times the gains from TP(Y) on Kinky Development ("extreme poverty") (smaller arrow) is larger in benefit-cost ratio than any other. The opposing claim, that research on national development is superior, is *either* that (a) a claim that the impact of non-RCT research on national development on national development outcomes (dashed arrow) times the impact of ND on a kinky aggregate development (e.g. "extreme poverty) (big arrow) is bigger *or* (b) that the impact non-RCT research on national development outcomes (dashed arrow) times the impact of national development outcomes on (inequality adjusted) aggregate human well-being is bigger in valuation term (given any reasonable valuation) than that of RCT on TP(Y) on extreme poverty.

There are two elements of Figures 2.7 and 2.8 that nearly all economists agree on.

First, the magnitude in dollar terms of gain from national development and sector-wide reforms are orders and orders of magnitude larger than possible with targeted programs. The *randomistas* do not typically argue that the gains to poverty from growth would not be large as, given the figures above, this is so obviously false, but rather argue the impact of research on growth is very small/weak/zero.

Second, the impact of RCT research on national development or sector-wide reforms is almost certainly limited. A reason I stressed that the processes of national development operate at an ontological level higher than the individual is that RCTs are typically only possible (and certainly only possible to "power up") when a large number of units can be assigned to "treatment" and "control" status. This is impossible for economy-wide or national-politics-wide or organization-wide phenomena.

Therefore, the most common claims by the sophisticated advocates of RCTs are some sets of the following:

- While the impact of national development on all types of well-being indicators is large and national development is sufficient for achieving kinky goals, and necessary for high goals, the impact of research on national development is very, very near zero (dashed arrows from non-RCT research to national development essentially don't exist) therefore even if the impact of RCT on actual targeted programs (for income Y or specific indicators S) is small, and only on the kinky, the valuation of the research is cost-effective if only because it is effective at all whereas the other types of research have (near) zero effectiveness.

Or,

- A different line of argument is that the valuation of human well-being is exclusively kinky so gains above the threshold don't matter therefore the national development impact on the non-kinky has very low value.

With either of the above claims, one has to add, *and:*

- RCTs are able to generate reliable and useful knowledge about targeted programs for income or specific goals.
- The reliable and useful knowledge generated by RCTs has to actually change the course of events, that is, the knowledge generated by RCTs has to be a (key) binding constraint to the scale of use of better targeted programs.

2.2.1 Widely Accepted Claim I: The Magnitudes of Gains from National Development Are Orders of Magnitude Larger than from Targeted Programs

Kenny and Pritchett (2013) show that, on basically any measure of human well-being progress in national development (called "drive") or gains in sector-wide efficacy (called "shift") dominate, by order of magnitude the gains from targeted programs (called "kink").

Pritchett, Sen, Kar, and Raihan (2016) estimate the net present value of GDP added (or lost) relative to a "business as usual" counter-factual from various episodes of growth or contraction. Our technical method of giving dates and sizes to growth episodes suggests that the growth accelerations in China in 1977 and 1991 produced NPV (net present value) gains of 2.65 trillion and 11.8 trillion (over 14 trillion in total). The growth accelerations in India in 1993 and 2002 produced gains of 1.1 trillion and 2.5 trillion (total of 3.6 trillion). Indonesia's growth acceleration in 1967 produced a NPV gain over BAU (business as usual) of 1.1 trillion. The absolute gains from Vietnam's acceleration in 1989 were smaller, $455 billion, but this was an NPV gain of $6,911 per capita. These growth episodes were also associated with a rapid reduction in "extreme poverty" to very low levels (Figure 2.5). The losses from decelerations relative to the BAU growth rate are also similarly massive. Brazil's loss from the 1980 deceleration episode was 7.5 trillion dollars, the loss to Indonesia from the 1996 East Asia crisis was near a trillion dollars, and the combined losses from the Mexico decelerations of 1981 and 1989 were 1.5 trillion. Many African countries, though small in absolute terms, had massive losses of NPV per capita from growth decelerations: Malawi 1978 P$9,600; Kenya 1967 P$13,300, Cote d Ivoire 1978 P$15,200.

The "livestock" trial published in science showed a complex, multi-faceted "graduation" approach to the ultra-poor raised year 3 incomes in 5 of 6 study sites. The magnitudes, averaged across the five sites, were that $4545 per household in costs in year 1 and 2 produced *$344 per household gain* or, on the assumption of a typical household size of 4, *$86 per person*. On the *assumption* that this year 3 amount persists forever, this implies, at a 5 percent discount rate, an average household NPV gross gain of $8472 in gains per household, which was about a 7 percent rate of return. Assuming crudely four people per HH that implies that an $1136

investment per person produces a once off level gain in year 4 of $86. Suppose we wanted to use the knowledge from this "gold standard" evaluation of an anti-poverty program to raise income in Vietnam by an NPV of $6,911. That would cost $333 billion dollars in program investments—more than Vietnam's *current* (post growth) total GDP or about three times *total* global development assistance.

The gains from well-functioning financial systems—especially from avoiding a large crisis—are huge. Estimates of the losses in 2014 to OECD GDP from the 2008 financial crisis were about 3.5 percent or 1.9 *trillion* dollars, if that is a "permanent" loss relative to a no-crisis counter-factual the Net Present Value of that (at 5 percent) is 38.2 trillion dollars. The US Federal Reserve estimates the NPV of the loss to the USA at US$70,000 for each citizen. The *total* stock of microfinance assets in 2016 was about 102 billion dollars. Suppose, at the wildest possible positive view, the annual gain to borrowers was 10 percent of the stock and this implies a gain to borrowers of 10.2 billion dollars. Suppose, again at the far reaches of optimism, rigorous research could somehow *double* that gain (relative to a counter-factual) then the gain would be an additional 10.2 billion dollars globally to microfinance borrowers. The losses from a single (large) global financial crisis were of the order of 200 times larger than the gains from doubling the total benefits from microfinance.

Raising the learning levels in basic education of children to prepare them for their twenty-first-century lives is hugely important. If one takes a view of the challenge how important is research on the enrollment impacts of conditional cash transfers? Using a recent assessment of learning in Zambia, the PISA-D, I estimated that, of the 360,000 children aged 15 in Zambia only 36 percent were in school and assessed and of those only an estimated *total of five children* (not 5 percent, five children, like the five fingers on your hand) could read at global high proficiency levels (PISA levels 4 or above, achieved by roughly a third of OECD students). Moreover, even if, through whatever heroic efforts, including say, expanded use of conditional cash transfers, enrollment of 15 year olds increased to 100 percent, at current levels of learning this would add only *14 children* who could read at global high proficiency levels. But Vietnam has learning performance that is massively better than Zambia's in ways that are not accounted for by targeted programs but rather appear to be the superior operation of a sector-wide education system.

2.2.2 Widely Accepted Claim II: RCT Studies Do Not Address National Development

Pritchett (2014a) draws on the Vivalt (forthcoming) review of RCT results to compare the topics on which enough RCTs have been done to compare results with some simple questions about whether topic X is even plausibly a major cause of growth. None of the common domains of RCTs (conditional cash transfers,

microfinance, improved cook stoves, deworming) are plausibly important determinants of the level of income or of growth. Nor do their advocates make that claim. The reason I emphasized the social nature of the four-fold national development transformation is that what the RCT needs to be successful as a research strategy is (a) (reasonably) clean assignment of units to "treatment" and "control" and (b) enough units for adequate statistical power. This is why, almost necessarily, the method lends itself well to *individualizable* (or small unit, like clinic or school or police station) interventions and not to studying the impact of policy on market performance or the evolution of the governance of a polity or the social transformation. Even if an RCT were to address these topics (like a study on information and voter behavior) they would do so in a way that, if and when the results were extrapolated to the scale of the relevant, they would have no more "rigor" or warrant as evidence than any other method as, in order to use the method precisely, the "general equilibrium" effects at the system scale had to be bracketed.

2.2.3 Needed But False Claim I: The Impact of Any Research (RCT or Otherwise) on National Development (or Sector-Wide Reforms) Is Vanishingly Small

Given the relative magnitudes of the gains to human well-being from national development and that the RCT method is not well applied to promoting national development or sector-wide changes, the argument has to be that *national development, including economic growth, is roughly impervious to any sort of research.*

This argument is at odds with commonly accepted interpretations of events in a number of countries. One, there are a number of countries (e.g. China, India, Vietnam, Indonesia) that said (1) "Based on our reading of the existing evidence (including from economists) we are going to shift from policy stance X to policy stance Y in order to accelerate growth," (2) these countries did in fact shift from policy stance X to Y, and (3) the countries did in fact have a large (to massive) acceleration of growth relative to BAU as measured by standard methods (Pritchett et al. 2016). One had to be particularly stubborn and clever to successfully make the argument: "Politicians changed policies to promote growth based on evidence and then there was a growth acceleration but (a) this was just dumb luck, the policy shift did not actually cause the shift in growth, something else did or (b) (more subtly) the adopted policies did work but that they did was just dumb luck as there was not enough evidence the policies would work for this to count as a win for 'evidence' changing policy."

There are also a fairly large number of countries that did the opposite. Economists (from their country and others) have said to the leadership of countries: (1) "If you persist in policy stance X you are going to experience large (to

massive) negative consequences for economic growth," (2) the leaders have not listened, and (3) there have been precisely the predicted negative consequences. The Venezuelan economy in 2018 was not spiraling into hyperinflation and in the midst of an economic depression because "economists have little useful to say about economic growth" in the sense their advice, if followed, would not be useful. If the argument is that research can produce reliable advice but this doesn't mean it will change the course of events, then the question is not whether it always works, but whether it *never* works to change the course of events. There are also cases in which governments who have said "based on what economists say we are going to switch paths to avoid massive downturns/hyperinflation," have done so, and it has worked (in the sense at least that a crisis did not happen). While the "growth accelerations" might have been hard to predict with standard policies (Hausmann, Pritchett, and Rodrik 2005) there is empirical evidence that "growth collapses" are rather more predictable (Breuer and McDermott 2013).

This is not to say that all research based claims about policies for growth have been right. The "lost decades" in Latin America and the "transition depression" in some (not all) former Soviet dominated countries are both examples of adopting policies for growth based on economists' recommendations that seemed not to work. However, as a chapter in this volume points out, among the top ten most prescribed medicines many work on only a third of the patients. So, just because a recommendation is not universally successful does not mean it is not a good recommendation. If I can give you a tip that increases your odds of winning a million dollar lottery by 10 percent, it is massively worthwhile. Moreover, recent reviews suggest the "pox on all the houses of growth research" stance and a view recommendations had been worthless are too extreme (e.g. Easterly 2019 on the "Washington Consensus," Irwin 2019 on trade).

Keep in mind from Table 2.1 just how small the expected effect of research on growth has to be as poverty reducing as what can be expected from improved poverty programs. Suppose that growth advice was given to 10 countries and in 9 of 10 it either was not adopted or was adopted and did not work but in 1 of 10 growth accelerated by 2 ppa for 20 years. Then even at this lack of efficacy it is still, for the poorest countries, poverty reducing. (And obviously if those countries that happen to adopt are large countries (China, India, Vietnam, Indonesia (1960s)) then the total well-being gain is massive even if it is mostly ineffective.)

Moreover, the weak performance of growth recommendations in the 1980s and 1990s could just as easily lead to recommendations for much *more* research on how to promote national development rather than less, given the value of getting good rather than bad advice on these hugely consequential issues. It is not as if economics was complacent and either ignoring the negative growth experiences from many episodes of policy reform (e.g. World Bank 2005) or sticking to "mindless growth regressions." An approach taking into account the episodic nature of developing country growth (e.g. Ben-David and Papell 1998, Pritchett 2000, Jones

and Olken 2008, Berg, Ostry, and Zettelmeyer 2012) married with a diagnostic approach (e.g. Hausmann, Rodrik, and Velasco 2008, Hausmann, Bailey, and Warner 2008, Rodrik 2009) was maturing even as the *randomista* movement was taking off.[16]

2.2.4 Needed but False Claim II: Valuation of Human Well-being Is "Kinky"

The other path in Figure 2.7 and Figure 2.8 into a priority within development field intellectual activity for RCTs is to adopt exclusively kinky measures of human well-being. This can make the fact that national development is a necessary condition for moderate to high levels of well-being and the massive gains from national development less compelling. I have written extensively elsewhere about why kinky goals generally, and low-bar poverty specifically, are illegitimate in every way: economically (Pritchett 2006, Pritchett 2013a), morally (Pritchett 2014c), politically (Gelbach and Pritchett 2002, Pritchett 2005, Pritchett 2014a, Pritchett 2014b) or as goals for development (Pritchett 2015), or development organizations (Pritchett 2013a), and so can be brief. The simple, but compelling, argument against kinky goals in either income or in specific indicators (e.g. water, education) is "introspection plus the Golden Rule."

Introspection. The essence of "kinky" is that gains to well-being are *exactly* zero above a low threshold. Ask yourself about yourself: did your personal valuation of income fall to exactly zero when your income passed some low threshold? Did your willingness to pay for higher quality sanitation facilities drop to zero at exactly an outdoor latrine? Did your personal valuation of education drop to zero when you finished primary school? The only honest answer is no.

Golden rule. A widespread (if not universal) principle of "moral realism" is something like the "golden rule"[17] (do unto others) or the Kantian categorical imperative ("Act only according to that maxim whereby you can, at the same time, will that it should become a universal law" (Kant 1785 (1998)). By the Golden Rule/Kantian Moral Imperative—and frankly common sense—adopting

[16] And siphoning off from growth research even funding intended to be channeled to growth research. For instance, the Crépon et al. (2015) paper re-reviewed Florent Bédécarrats, Isabelle Guérin, Solène Morvant-Roux, and François Roubaud (2019a) and discussed in this volume (Chapter 7) was funded and promoted by the International Growth Centre, which was originally funded by DFID to improve "growth analytics" in order to lead to more prioritized and pragmatic recommendations to countries for policies to promote growth. Whatever the paper's (de)merits substantively it is a paper about a targeted program and no one can even pretend it is a paper about promoting national development or growth.

[17] Parfit (2011) argues that three common approaches to moral questions— the Kantian deontological, consequentialism, and contractualism—ultimately converge to the same answers and that these are "correct" answers.

for the general assessment of the well-being of other people by a standard you would never accept for yourself is morally wrong.

Any attempt to "solve" this by claiming the objective function is a "combination" of kinky and non-kinky goals means the overall goals are non-kinky and it is just a question of weights but the massive gains above the threshold are relevant. The replacement of the kinky MDGs with the broad and expansive SDGs should have ended the relevance of the kinky as legitimate expression of development (Pritchett 2015).

2.2.5 Needed but False Claim III: RCTs Can Reliably Generate Evidence that Improves Targeted Programs Aimed at Kinky (Aggregate or Specific) Development Goals

Another path to claiming RCT studies as the "best investment" is to claim that impact evaluation of programs/projects using RCTs are likely to produce rigorous, reliable, and usable evidence that can lead to the design of more effective programs. As I, and many others, including many authors in this volume, have argued: (a) these claims never had any solid evidence but were just asserted on faith, (b) claims that RCTs would "resolve debates" about impacts based on heterogeneity in observational studies were ex ante not just empirically unlikely but logically impossible (Pritchett and Sandefur 2013a), (c) empirically the reviews of empirical studies fail to show sufficient consistency to be reliable (Vivalt forthcoming), even within specific topics like improving learning in basic education (Evans and Popova 2016) or deworming (e.g. the "Worm Wars"), and the variability across "rigorous" studies is sufficient that, at least in some instances, relying on the "rigorous" evidence would not reduce the prediction error about program impact in a given context relative to simple methods (Pritchett and Sandefur 2015)—which is exactly what everyone except the *randomistas* expected (Pritchett 2018c), (d) the "construct validity" (the robustness of results across variations in the design space) of RCTs is low (Nadel and Pritchett 2016, Kerwin and Thornton 2018, Kaffenberger 2018), and (e) one cannot use results "proven" with one implementer to extrapolate to impact when implemented by another organization, particularly from an NGO "proof of concept" to scaling with government (Bold et al. 2018, Vivalt forthcoming).

The "livestock study" (Banerjee et al. 2015a) mentioned by Professor Blattman is sometimes taken as the "proof" than one can create "gold standard" evidence that could guide effective anti-poverty programming. In that context, there are seven points worth nothing. First, the IRR is 7 percent, which is not particularly impressive; it would not pass the 10 percent rate of return traditionally used World Bank project cost benefit analysis. Second, in my calculations above, I was being generous and not including in these calculations one of the six countries, Honduras, in

which the livestock (chickens!) died and hence the program had pretty substantial negative impacts on households so the average given does not reflect all experiences. Third, it is not clear the program beats a cash transfer as the costs to produce the gains are very high. Fourth, there is not (yet) "rigorous" evidence that the gains of the program will be sustained. Their calculations suggesting this program produces positive NPV requires the *assumption* that the year three gains are sustained into the distance future, an assumption not supported by their data. If one uses the observed fall in measured annual durables consumption from year 2 to year 3 and extrapolates future income streams using that decay the NPV is *negative* for all but two of the six countries. Fifth, Bauchet, Mordouch, and Ravi (2015) did an impact evaluation of a very similar program in South India and they find no impact on income or assets, they argue because the local economy was growing robustly so the livestock option was not attractive so we already know *for sure* these results lack external validity, at least across some external conditions. Sixth, one suspects there is a lack of "construct validity" in the sense that this "multi-faceted" program was complex and had many elements in part because the design was the result of a long period of more informal "trial and error" and "experiential learning" (Pritchett et al. 2012) by BRAC and hence even minor variants in the design or the fidelity of its implementation might not produce the positive results. Seventh, while the study was done across multiple sites, responsibility for implementation was the responsibility of the same organization in all sites, so the robustness of these results to any other organization is not at all assured.

The relevance for this chapter is that if one wants to claim that the "best investment" is research into a topic that has very, very, limited upside gains (e.g. design of sector specific targeted programs) compared to other research that has massive upside gains (e.g. promotion of national development) the offsetting gains in likelihood of producing reliable, usable results have to be very large. If research into national development has a one in a thousand chance of producing usable results and RCTs a 100 percent chance this is a powerful argument in favor of (some) RCT research. However, there is no compelling or persuasive evidence or argument that the likelihood of producing reliable and useable results from a given magnitude of effort into RCTs is *higher at all*, much less that it is orders of magnitude higher.

2.2.6 Needed but (Probably) False Claim IV: Knowledge of the Type RCTs Can Generate Is a Binding Constraint to the Adoption and Implementation of Better Targeted Programs

In order for a proposed public policy/program/project to have (sustained) impact it has to meet a "trinity": it has to be (1) "technically correct" (if implemented it has to be based on a correct set of causal claims about links from inputs to

activities to outputs to outcomes), (2) "politically supportable" (one has to be able to generate and sustain a political coalition with sufficient power to authorize the needed actions and resources), and (3) "administratively feasible" (one has to be able, with available administrative capability (or the capability that can be mobilized or created) to implement the program with sufficient fidelity to achieve the outcomes). Claims about improvements in human well-being from knowledge gained from RCTs depend on claims that knowledge about program design of the type RCTs can generate are "the" (or at least "a") binding constraint versus other constraints on effective action (Pritchett 2018b). But it is not obvious policy design matters for outcomes. Chong et al. (2012) show that, for a very specific policy outcome measure, return of misaddressed foreign mail, (a) the de jure policy is exactly the same in all countries and (b) the outcome, percent of mail return in compliance with the de jure policy, the outcome varies by as much as it possibly can (zero percent to 100 percent) and hence (c) all of the variation is due to implementation, none to policy.

The "design space" for a project/program aimed at any objective (e.g. women's empowerment, reducing farmer income variability, increasing savings, reducing morbidity from water-borne diseases, etc.) is likely to be large and complex (and unknown) in that there are many choices (e.g. who is responsible for what actions, how frequent should visits be, what is the content of informational messages transmitted, what is the magnitude of a loan, etc.) and many possibilities for each choice and some elements crucial for success might not even be known at the design stage. Doing an RCT establishes an estimate of "impact," which is a point (or set of points, one for each treatment arm) on the "response surface" of outputs or outcomes over a particular design. The previous section (Section 2.2.5) was about how useful this inference about a point or set of points is when the response surface could vary across contexts or be very rugged (non-robust) with respect to design. But there is an additional concern that knowing the response surface over a project/program design that is administratively or politically impossible has limited or zero value[18] (Gass and Pritchett 2017, Pritchett 2018a).

Knowing that projects/programs would have impact X or Y or Z if adopted in contexts where, even when X or Y or Z are fully known and agreed, these projects/programs have zero probability of political adoption may contribute to disciplinary knowledge but cannot be claimed to have benefits for human well-being. Pritchett (2010b), drawing on Filmer and Pritchett (Filmer and Pritchett 1999), argues that much of the advocacy around the usefulness of RCTs for "policy-making" presumes a "normative as positive" model of politics, even in domains in which that model has been shown to be demonstrably false. One doesn't have to buy wholesale into public choice theory to accept that one cannot take seriously as a positive model

[18] This is just the obvious Kuhn-Tucker point from optimization subject to (potentially) many constraints; the "Lagrangian" or "shadow price" on slack constraints is zero.

the idea that actors in the public sector (politicians, policy-makers, senior technocrats) are optimizing a social welfare function and constrained only by their knowledge of "what works."

The same logic is true of the capability of organizations expected to implement programs. Knowing that program would have impact X *if* it could be implemented with fidelity doesn't mean the existing organizations in the country, public or private, have that capability. A fair number of existing RCTs have not been able to demonstrate the causal link between the design of the intervention, outputs of the implementing organization, and outcomes. Rather, what the experiment learned was that, even in the limited context of an experiment, the "treatment" (whether it was pay for performance, citizen information or top-down instructions) could not alter the relevant behavior of the implementing agents to produce "outputs," e.g. Banerjee, Duflo, and Glennerster 2008 (assistant nurse midwives in Rajasthan), Banerjee et al. 2010 (public school headmasters, teachers in Uttar Pradesh), Banerjee et al. (2012) (police in Rajasthan). And examples where an experiment worked to produce outcomes when implemented with an NGO did not work when scaled by the government (e.g. compare Duflo, Hanna, and Ryan (2012) on cameras in classrooms in NGO schools to Dhaliwal and Hanna (2013) on biometrics in public sector health clinics in Karnataka). How much of observed variation in poverty or sector programs across countries is due to the large differences across countries in implementation capability rather than policy design (the large variation in country indicators of state capability and their connection with measured human well-being outcomes across countries is in Table 2.2).

There are pretty good arguments that the "technical" or "codifiable" knowledge that RCTs are best placed to produce are, at best, a minor constraint on the adoption and effective implementation of targeted programs (Pritchett 2018a, 2018b) versus political constraints on the "want to" and the capability of "can do" and neither of these are affected by the results of RCTs. In contrast, a good argument can be made that the use of existing knowledge in a given country is endogenous to politics and capability, rather than an exogenous factor, as the "codifiable" part of knowledge is a public good that, being non-rival and non-excludable, should diffuse quickly and easily.

2.3 Conclusion

An impact evaluation with an RCT seems to be not really a tool for countries and their governments or for agencies interested in promoting development at all. Rather, it mostly seems a tool to guide that small part of the development process that is "charity" or "philanthropic" that is (a) going to give relative small amounts of money, (b) will not or cannot work though national (or state or local) governments, (c) has relatively "kinky" valuations (perhaps in part because they are

rationing tiny resources), and (d) care about the ability of being able to attribute the gain in well-being causally to their specific intervention (rather than about indirect effects). Charity work is a good thing and if charity work can be done better guided by evidence from RCTs that is a good thing. A focus on charity work is likely how Bill Gates and Chris Blattman get to talking about chickens and their impact.

However, to confuse this tiny little segment of the world with the broader process of development is not remotely plausible. South Korea today is not the South Korea of the early 1960s because its government did a better job promoting ownership of chickens. The world today is night and day better on nearly all objective measures of human well-being because of broad-based national development (including economic growth) and improved sector-wide performance of the kind development was meant to promote (Pritchett 2017). To imagine that the tools that international NGOs want to use to identify effective humanitarian interventions for the poorest of the poor, particularly when those interventions considered are limited to those where impacts can be directly attributable to the NGO's actions, are also the "best investment" in poverty reduction, much less the best investment in development, is not a considered promotion of a method but madness.

3
The Disruptive Power of RCTs

Jonathan Morduch

3.1 Introduction

There are two distinct ways that RCTs are used in development economics. In the first, RCTs are used to measure impact. In the second, RCTs are used to explore the nature of economic contracts, behaviors, and institutions. The two kinds of RCT are often lumped together by critics, but the two strands speak to very different questions and serve different purposes. Understanding the power of RCTs, and disentangling debates around RCTs, requires first separating the two modes.

Critics are especially uncomfortable with elevating RCTs as the favored tool for evaluation, but one can accept their criticisms—in whole or in part—and still embrace the importance of RCTs (and want to encourage far more RCTs) in the cause of experimentation. Should the randomistas rule (Ravallion, Chapter 1, this volume)? No. Are RCTs a gold standard (Bédécarrats, Guérin, and Roubaud 2019)? No. In practice, however, RCTs have been—and will continue to be—particularly useful exploratory tools.

The first use for RCTs (and the focus of the heaviest criticism) is the promotion of impact evaluation through randomized methods. The criticism is less often about RCTs per se than about putting them on a pedestal, with a special status that accords them more credibility than other evaluation methods. These RCTs focus on evaluating government or NGO programs and policies, and the hope of proponents is that having more credible measures of impact through randomization will mean better investments and interventions (Glennerster and Takavarasha 2013; Kremer 2003, Banerjee and Duflo 2009). The questions usually focus on "what works." Older studies include RCTs of government programs like Mexico's Progresa conditional cash transfer program (Levy 2006) and the US Job Training Partnership Act (Lalonde 1986), and, in the most recent wave, evaluations of NGO programs like the microcredit RCTs rounded up in Banerjee, Karlan, and Zinman (2015). For the most part, researchers design the evaluations but not the interventions. Much of the debate in this book tackles whether and how such RCT evidence should matter.

The second kind of RCT has a different character. It aligns with the experimental mindset increasingly adopted by development economists, with RCTs as a critical

Jonathan Morduch, *The Disruptive Power of RCTs* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0005

methodological innovation. While some economic experiments involve lab-based hypothetical scenarios (e.g., Davis and Holt 1993), this strand of RCTs involves experiments in real settings. The studies are based on experimentally controlled manipulations of price structures, contracts, teaching methods, healthcare protocols, bureaucratic processes, and the like. Here, researchers participate actively in the design of the actual programs and policies, usually together with a government agency, business, or NGO. The questions asked are exploratory, theory-driven, and motivated by the desire to understand economic possibilities and constraints. The contexts are often limited-scale pilots or limited-time trials. The questions are less often about "what works" than "how and why?" or "what could be?" While RCTs for evaluation are criticized for saying little about "why"—why impact is small or large or appears for some people but not others—these studies center on explanation. They ultimately ask whether the world works in the ways that economic theory says it should. The power of these RCTs lies in how they disrupt business-as-usual by manipulating economic environments and thereby allowing vision into what would otherwise remain unseen or untried.

The line between the two kinds of RCTs can be fuzzy, blurred by both RCT advocates and critics, and the aim of this chapter is to clarify the modes and illustrate the experimental mindset in development economics. The view I put forward is not necessarily what would be made by a full-throated randomista, but it aligns with how RCTs are often used in practice.[1]

The first section of this chapter describes the rise of the experimental mindset coupled with RCTs. The second section gives three examples of RCTs that probe questions related to prices, contracts, and the use of financial services in poor communities. The third locates the focus of RCTs on poverty-reducing interventions and the provision of private goods, and considers the argument that RCTs push focus away from studying the systemic forces that shape economies.

3.2 Expanding Knowledge by Creating Variation

The primacy accorded to RCTs for evaluation—along with related methods like natural experiments and regression discontinuity designs—leads to the fear that the rise of RCTs for evaluation unduly and unhelpfully downgrades other ways of assessing what works (e.g., linear regression, conventional instrumental variables, ethnography and qualitative evaluation, and machine learning with Big Data). More worrying, giving primacy to these kinds of RCTs risks restricting attention to the set of economic interventions that are most amenable to randomized trials.

[1] See Ogden (2017) for a view from academics and practitioners engaged with RCTs, with a theme around differing theories of change.

The fear, at the extreme, is that giving RCTs a special status in determining "what works" could lead to a loss of knowledge, especially relative to what learning from a diversity of approaches could deliver (Ruhm 2019).

Detractors also worry that advocates exaggerate the precision and the ease of generalizability of RCTs (Deaton and Cartwright 2018). They worry that the kind of evaluative information generated by RCTs is often of limited political and practical value (Drèze 2018b, Pritchett 2014c), and is vulnerable to misinterpretation for lack of context (Morvant-Roux et al. 2014). Like other evaluation methods, RCTs have difficulty providing crisp answers, especially when, as is often the case, it is necessary to extrapolate from a study in one place to a policy environment in another (Cartwright and Hardie 2012, Pritchett and Sandefur 2015, Bisbee et al. 2017).[2]

Perhaps most worrying, critics argue that the interventions most amenable to evaluation by RCTs are too small, too limited, and too particular. Within economics, RCTs are an easier fit for studies involving private goods than public goods. Moreover, RCTs often focus on marginal impact and on impact on marginal subpopulations (Wydick 2016). They can be used to measure short-term impact when microcredit enters a new region, for example, but not to evaluate how the original customers have fared since the microcredit organization's start (Cull and Morduch 2018).

From a broader vantage, by focusing on small steps to improve the implementation of existing ideas, evidence of impact from RCTs tends to only speak indirectly about the broader structures that perpetuate poverty and inequality. By this view, giving complete primacy to RCTs for evaluation would restrict admissible evidence on "what works" and ultimately narrow understandings of complex economic and social phenomena (Bédécarrats, Guérin, and Roubaud 2019).

In contrast, the RCTs for exploration (the second kind of RCT above) more clearly expand knowledge, and most RCTs published by development economists take this direction. The experimental mindset responds to the fact that key variables may not move much in the natural course of things, so experiments are needed to create relevant variation. Prices may not change much in a given moment or sample, nor contracts. Governments, clinics, schools may all act uniformly in a given range. The result is that, while researchers can explore theoretical predictions, they have little hope to take them to the data. Without experimentation, there is too little to observe and thus too little to analyze.

These exploratory RCTs have limits too: it is tempting to draw overly-strong policy conclusions from the trials and pilots, rather than taking them for what they are: informative and provocative but contingent. Yet, at the same time, criticizing these RCTs for being pilots or trials risks missing how they can aggregate to create sharper, more expansive understandings of constraints and possibilities.

[2] As Imbens (2018) notes, however, scholars using RCTs are aware of the limits and are responding with expanded approaches (e.g., Bates and Glennerster 2017).

Although Angus Deaton and Nancy Cartwright argue against giving evidence from RCTs a special status, they note,

> RCTs are often convenient ways to introduce experimentally-controlled variance—if you want to see what happens, then kick it and see, twist the lion's tail . . . (Deaton and Cartwright 2018, 17).

From the perspective of economic knowledge, twisting the lion's tail with the help of RCTs has pushed researchers to better understand economic theory and question assumptions that were once considered settled.

Consider the case of crop insurance, a product with much potential given the risks of rain-fed agriculture. In practice, however, crop insurance (and its newer variant: index-based rainfall insurance) has been particularly difficult product to sell to farmers. Casaburi and Willis (2018), for example, show that only 5 percent of Kenyan sugar-cane farmers in their sample purchased rainfall insurance, a finding that reinforces the sense that potential customers are wary of these products, might not understand or trust them, are content to rely on informal mechanisms, and/or find the products too poorly designed or too expensive. Casaburi and Willis, however, use an RCT to experiment with the timing of when the insurance is sold. They ask whether the problem is not mainly the price nor the understanding of customers. Instead, could the low take-up rate occur because insurers ask for the premium to be paid in a lump sum before the planting season, a time when most money is being invested in crops? By randomizing the timing of payment, pushing it to harvest-time (when farmers have liquidity) for a sample of customers, they show an increase in the take-up rate to 72 percent. In contrast, reducing the cost of the insurance by 30 percent (but not delaying the timing of payment) only increased demand by one percentage point. The RCT allowed everything else to be kept the same, and, while the finding is not revolutionary, it helps expand perceptions of the problem. Whether the exact parameter is transportable or not is less important than that the study highlights timing and liquidity as constraints to insurance demand to consider seriously in other settings (in addition to highlighting a practical response to the problem).[3]

Casaburi and Willis's experiment in Kenya informs the work of Belissa et al. (2019) in Ethiopia. They too investigate the role of liquidity on the take-up of insurance, again asking whether demand is greater when farmers can pay after the harvest when liquidity is greater. They additionally explore the role of promoting insurance through *Iddirs*, local informal risk-sharing mechanisms

[3] Similarly, Jonathan Bauchet and I investigate the demand for a life insurance product in Mexico sold to poor women. Using a natural experiment, we find that demand rises by over 59% when customers are allowed to pay in small weekly installments rather than in an upfront sum (Bauchet and Morduch 2019).

used by farmers. The Belissa et al. (2019) design involves 8579 individuals and 144 *Iddirs*. The RCT has six treatment arms. The first is a control group that is offered a standard index-based rainfall insurance contract that requires payment before the insurance takes effect. The second group is similar but the product is promoted by a local leader. The third group is also like the first, but delayed payments are allowed. The fourth is similar to the third, but the purchaser is asked to formally sign a binding contract committing to pay the premium after the harvest. The fifth group gets the insurance product promoted through the Iddir (with the possibility of delayed payment), and the sixth gets everything—the possibility of delayed payment, the requirement to sign a binding contract, and promotion of insurance through the Iddir.

Although less dramatic than in the Casaburi and Willis study, delaying the timing of the payment turns out to be substantial for the farmers in Ethiopia, increasing take-up from 8 percent to 24 percent. Combining the delayed payment with promotion through the *Iddirs* intensifies the impact, bringing take-up rates to 43 percent. Promoting insurance via *Iddirs* not only helps bring credibility to the insurance product, it also facilitates the collective purchase of insurance against an explicit background of informal insurance. The study, though, shows that about 15 percent of farmers who agreed to pay after harvest in fact defaulted on their commitments to pay, a level high enough to threaten the economic viability of the insurance product.

The two insurance studies illustrate the fundamental distinction between RCTs for exploration—researcher-designed experiments that open the box to probe mechanisms—versus RCTs for evaluating the impact of established programs. Neither study here measures the impact of insurance on farmers. The main aim is not to evaluate whether insurance "works," and, in line with that, neither study has a pure control group with no intervention. Instead, in both studies the control group has the chance to buy a standard insurance product. Both studies then explore what happens when the products are redesigned in systematic ways to gauge farmer behavior and the viability of the products. The specific results of neither experiment can be extrapolated to other contexts, but the nature of the innovations (the delayed timing of payments, marketing through local groups) and broad concerns (illiquidity, the risk of post-harvest default) can be.

When it comes to impact evaluation, RCTs are often promoted for reducing selection bias due to nonrandom program access, but the two insurance examples show that selection bias is just one of several big challenges in empirical development economics. Here, a main problem is the lack of relevant variation in insurance contracts (especially the lack of observed contracts offering post-harvest payments), a problem exposed via experimentation through the RCT. Neither study had to be an RCT, but both had to involve experimentation and product redesign. Both had to "twist the lion's tale." The fact that both sets of researchers chose to use RCTs stems from the practicality of joining experimentation with randomization in an exploratory mode.

While Ravallion (Chapter 1, this volume) traces the history of RCTs in economics to experiments in the 1950s and 1960s (see also Gueron 2017), the notable rise of RCTs in development economics started in the 1990s, following a period of methodological ferment that, among other outcomes, led to focuses on natural experiments (Angrist and Krueger 1999). The move from natural experiments to RCTs was not a large one conceptually, pioneered by Harvard's Michael Kremer in Kenya, and solidified later by the establishment of MIT's J-PAL (Kremer 2003, Banerjee and Duflo 2009; see Ogden 2017 for descriptions of process and motivations from Banerjee, Duflo, and Kremer). Kremer and his colleagues took part in designing the interventions, unlike the previous evaluation-based RCTs that tested government-designed interventions. Kremer (2003) summarizes a series of early experiments to improve schooling outcomes in Kenya, including providing free breakfasts, supplying school uniforms, adding textbooks, de-worming children, and introducing more teachers. Several of the interventions increased school participation substantially at relatively low cost.

The examples show where confusion arises about the types of RCTs. Kremer (2003) describes the RCTs as evaluations of the "what works" sort (in the sense above). Yet, without diminishing their value, they are in essence exploratory. They are largely pilot programs, not large-scale public programs. They usefully document possibilities and constraints, providing an important opening or next step rather than the last word.

3.3 The Ubiquity of Suboptimality and the Potential for Innovation

Deaton and Cartwright (2018) are careful to distinguish "what works" RCTs from exploratory "how and why" RCTs.[4] In this context, they consider "when RCTs speak for themselves" and situations with "no extrapolation or generalization required":

> For some things we want to learn, an RCT is enough by itself. An RCT may provide a counter-example to a general theoretical proposition, either to the proposition itself (a simple refutation test) or to some consequence of it (a complex refutation test). An RCT may also confirm a prediction of a theory, and although this does not confirm the theory, it is evidence in its favor, especially if the prediction seems inherently unlikely in advance.

What's at stake in most exploratory RCTs is seldom refuting theory in the sense of Deaton and Cartwright. The two insurance examples, for example, center on

[4] Deaton and Cartwright (2018) anchor a special issue of *Social Science and Medicine* focused on "Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue," edited by Ichiro Kawachi, S.V. Subramanian, and Ryan Mowat and featuring 19 responses from leading statisticians and social scientists.

well-known ideas (illiquidity, lack of trust), and their importance is unsurprising (in the sense that they are both likely to be somewhere on the list of challenges to providing insurance). Instead, what's ultimately at issue is how much faith to place in constrained optimization. A fundamental tenet of neoclassical economics is the idea that markets yield optimal institutions, goods and services, and prices. In theory, the disciplining function of the market should weed out suboptimal forms. This tenet holds even in second-best or third-best worlds with constraints like asymmetric information and imperfect contract enforcement (Stiglitz and Weiss 1981). In essence, modern economic theory says that what we see is not necessarily perfect, but it is as good as it can get.[5] In other words, existing insurance processes and products should already incorporate ways to deal with problems of liquidity and trust to the extent feasible.

But is that generally true? Muhammad Yunus's experimentation with credit contracts in the 1970s, which led to the development of microcredit, illustrates a case where tinkering and redesign created a genuine improvement over what the market had delivered. New contracts led loan default rates to drop sharply and profitability to become possible even when lenders charged relatively modest interest rates (Armendàriz and Morduch 2010). What economists thought had been a constrained-optimal outcome turned out not to be. And even Yunus's tinkering was not the last word in microfinance innovation (e.g., Rai and Sjöstrm 2004, Field et al. 2013).

The exploratory RCTs carry on in this spirit, driven by experimentation, helping to map how far existing institutions and choices are from what could be possible. Increasingly, the RCTs also map why innovation has not happened (for example, fear about the relatively high rate of default documented by Belissa et al.) and, often, test practical steps to mitigate problems. The contribution of exploratory RCTs is seldom a test of a specific theoretical proposition (like "are individuals rational?") but is a demonstration of an innovation or experimental manipulation that exposes (or deepens understandings of) suboptimality.

3.4 Why RCTs?

Writing about testing theory, Deaton and Cartwright note that generalizability is not always the major concern. They continue,

> [Theory-testing] is all familiar territory, and there is nothing unique about an RCT; it is simply one among many testing procedures (Deaton and Cartwright 2018, 12).

[5] A fundamental result in the economics of information is that equilibria may not even be constrained efficient (Stiglitz 1986). The RCT research program can be seen as showing a far wider range of circumstances with inefficient outcomes.

At a high level of generality, it must be true that "there is nothing unique about an RCT" here. There are, of course, other methods that can prod theory, demonstrate suboptimality, disrupt, surprise, and expand economic frameworks. Methodologists are now building on the experimental mindset, in some cases improving on bread-and-butter RCTs (e.g., Kasy and Sautman 2019) and in other cases integrating randomized assignment with ethnography (e.g., Duncan, Huston, and Weisner 2007). There are also non-randomized methods that can be used to analyze exogenously-created disruptions. But RCTs, coupled with an experimental mindset, have been particularly helpful in practice. Part of the case for RCTs when used for exploration echoes the case for using RCTs for evaluation: selection bias is a constant worry, and RCTs can help rein it in (while, admittedly, creating other issues). But another part of the case is that when you are *already* manipulating the economic environment in an experimental mode, randomization appears as a smaller stretch.

Researchers using RCTs wonder why one would want to use an alternative method to study their given question in their given place. Why study price elasticities for insecticide-treated bednets, for example, using a non-randomized approach when randomizing prices is feasible? In this line, Deaton and Cartwright (2018) note that a frequent response to their critique of RCTs is: "OK, you have highlighted some of the problems with RCTs, but other methods have all of those problems, plus problems of their own" (Deaton and Cartwright 2018, 16). Deaton and Cartwright refuse to accept that retort because they find that reliance on RCTs substitutes one set of problems for another set. As Ravallion (Chapter 1, this volume) notes, for example, problems in RCTs arise with selective non-compliance and "essential heterogeneity."

Still, the most prominent alternative approaches to causal inference (especially applications of instrumental variables) are subject to well-known limits. If nothing else, the history of empirical development economics has established that (1) selection bias often matters a lot and (2) plausible instrumental variables and natural experiments are hard to find. This is true across economics, but particularly so in development economics.

One illustration is offered by Beaman et al. (2018b), who construct an experiment to measure selection into borrowing in a sample of farmers in Mali. Their aim is to measure returns to capital and the impact of microcredit for agriculture, paying attention to the possibility that the most promising farmers are more likely to borrow than others. With no exogenous, excludable variation of prices and other external factors, estimation with instrumental variables is not feasible. So, instead, they construct a two-stage RCT. To get insight into the extent of selection bias, Beaman et al. randomly select 88 out of 198 villages in Mali in which to offer loans through a local microfinance provider. They then randomize the allocation of capital grants to a sample within the 110 villages without the loans and a sample of non-borrowers in the 88 villages that received the loans. They can then measure returns to capital for both borrowers and non-borrowers.

On average, returns to capital were large and positive with clear evidence of liquidity constraints. Recipients of the capital grants (which were worth about $140) in the 110 villages without the loans increased land under cultivation by 8 percent, use of fertilizer by 16 percent, and total input value by 15 percent. As a consequence, net revenue increased by 13 percent. Similar results were found for borrowers in the 88 villages that received microcredit (counter to well-publicized negative results summarized by Banerjee, Karlan, and Zinman 2015). But farmers who chose not to borrow but who had access to microcredit had essentially zero returns to capital at the margin. Thus, comparing the returns of borrowers to non-borrowers—without accounting for the endogeneity of borrowing and the heterogeneity of returns—would greatly overstate the net returns to microcredit access.

The lack of plausible instrumental variables tends to be greater in microeconomic studies of development because market failures drive interlinkages between household choices and between markets, especially in informal settings (e.g., Stiglitz 1986, Bardhan 1984). It is thus harder to find plausible excludable variables because, without complete markets, more elements of the economy prove to be endogenous. Empiricists working on the canonical agricultural household model (Singh, Squire, and Strauss 1986), for example, have exploited a recursive property that justified analysis of production independent of consumption variables, but the reverse does not hold, effectively ruling out the use of any production variables as instruments when analyzing consumption choices of producer-consumer households (including farmers and small-scale entrepreneurs). And even the recursiveness property depends on strong assumptions about the completeness of markets, including insurance markets.

The result is many good ideas but far fewer convincing ways to challenge and test the ideas—even when it is possible to observe naturally occurring variation in the economic environment. Moreover, using instrumental variables often leads to situations in which instruments may not be fully convincing but nonetheless estimated parameters are substantially affected by IV estimation.[6] The Local Average Treatment Effects (LATE) framework helps to explain why: with heterogeneous treatment effects, OLS and IV essentially estimate different parameters (Imbens and Angrist 1994, Imbens 2010). Like the understanding of weak instruments (Staiger and Stock 1997), the LATE framework challenged what can be learned from instrumental variables strategies. It became clear that differences between OLS and IV estimates could not be assumed to result solely (or even

[6] This is the case, for example, with Pitt and Khandker (1998), a well-known non-randomized evaluation of microcredit in Bangladesh that relied on the assumption of treatment effect homogeneity and the use of particular functional forms for identification, and which ultimately proved not to be robust even on its own terms. For a critical discussion, see Roodman and Morduch 2014.

mostly) from removing bias (a natural interpretation only under the assumption of homogeneous treatment effects). Instead, IV generates particular parameters that are specific to the interaction of the instrument and the endogenous variable when treatment effects are heterogeneous (Heckman and Urzua 2010). While RCTs also generate parameters that are local and specific, their interpretation can be read through the experimental design. They thus offer a mode of interpretation that is often clearer than in a typical LATE from an IV regression, especially one that does not draw on a natural experiment and one with multiple continuously defined instruments (Samii 2016).

3.5 Three Examples

To illustrate RCTs for exploration, I describe three examples of experiments with contracts, prices, and access to financial markets and products:

3.5.1 Microcredit contracts

The typical microcredit contract takes an unexpected form for a business loan. Although described as a loan for investment in small-scale enterprise, loans look more like consumer loans, with contracts that require repayment in regular installments starting shortly after the loan has been disbursed. In loans from Grameen Bank, for example, the installments are weekly and start the week after disbursement. In effect, the loan size is diminished since part of the loan must be returned to the lender nearly immediately. This structure, however, helps minimize the size of installments and has been promoted as a way to maintain high loan repayment rates (Armendàriz and Morduch 2010).

Might this structure, though, discourage investment and reduce profits for customers (and, possibly, local economic growth)? Might borrowers do better if they had more time to invest before repayments start? Field et al. (2013) designed an RCT to test that proposition, asking whether the "classic" microfinance contract inhibits investment in high-return business opportunities? They worked with an NGO that served with women in low-income neighborhoods of Kolkata, India. Each client received an individual-liability loan varying in from Rs. 4000 ($90) to Rs. 10,000 ($225), with a modal loan amount of Rs. 8000.

After group formation and loan approval (but prior to loan disbursement), groups were randomized into two contracts. In the control group, 85 groups were assigned to the regular debt contract with repayment in fixed installments starting two weeks after loan disbursement. In the treatment group, 84 groups were assigned to a contract that included a grace period of two months. Other features of the loan contract were held constant. The total interest paid was identical, and

once repayment began, all groups repaid every two weeks over 44 weeks, at a group meeting.[7]

Three years later, the new contract looked like a success: borrowers in the treatment group had 57 percent higher profit levels on average. They were also using 81 percent more capital and taking greater risks as they invested more. The problem, from the lender's vantage, however, was that repayment problems increased by three times: 52 weeks after the loan should have been fully repaid, 6 percent in the treatment group had not fully repaid compared to under 2 percent of the control group. The repayment problems were large enough that the contract was not profitable at feasible interest rates.

The study is not an impact study that asks: "does it work?" Instead, it investigates the nature of contracts and constraints, comparing one kind of contract against another. In the course of the study, a measure of returns to capital could be estimated (11–13 percent per month), suggesting that access to more capital would be welfare-enhancing, but that was not the main aim of the study. Instead, the RCT helps to get at a persistent question: why do the measured impact of microcredit appear to be so modest (Banerjee, Karlan, and Zinman 2015)? Are lending methods part of the problem? Can they be improved?

The RCT departs from market surveys by testing a real product rather than asking about preferences over hypothetical scenarios. A market survey might reveal a preference for delayed repayments, but would likely say little about the consequences for investment, business outcomes, and loan repayments. A trial could be run without an RCT, of course, but coupling with an RCT is a natural way to clarify comparisons.

3.5.2 Microcredit Interest Rates

The example above focuses on microcredit contracts, but the most important microcredit innovation was likely the choice to raise interest rates. It was not an obvious move. State-run banks had been created explicitly to provide subsidized credit in poor areas because it was thought that customers could not pay high interest rates. But early leaders in the field felt pressure to cover their basic costs, so interest revenue was imperative. With little more than casual evidence, micro-lenders reasoned that poor households seemed to borrow regularly from money-lenders who charged 5 percent or 10 percent per month, so charging 20 percent or 30 percent per year did not seem prohibitive. Using the logic of diminishing marginal returns to capital, micro-lenders also reasoned that capital-starved

[7] Because the treatment group had loans with a longer debt maturity (55 as opposed to 44 weeks before the full loan amount was due), they faced a slightly lower effective interest rate on the loan.

entrepreneurs could have high returns to their first increments of capital (Armendàriz and Morduch 2010).

The mantra soon became: "poor households need access to credit, not cheap credit." Implicit in this conclusion was the assumption that the elasticity of loan demand with respect to interest rates was effectively zero (Morduch 2000). Accordingly, interest rates were raised. Cull, Demirgüç-Kunt, and Morduch (2018), for example, show that for a sample of 1330 microfinance institutions between 2005 and 2009, average inflation-adjusted microfinance interest rates were 25 percent per year (21 percent at the median). These interest rates allowed microfinance institutions to reduce dependence on subsidy, although only about a quarter were truly free of subsidy.

Lenders assured themselves that there were limited tradeoffs with outreach. But the essential assumption—that the demand elasticity with respect to interest rates was zero—was untested and largely untestable. With available data, the challenges were: (1) lending institutions seldom changed interest rates, so there was little to analyze; (2) while different lenders charged different interest rates, so much else differed between institutions that separating out the causal impact of interest rates by comparing borrowing levels across institutions was a non-starter; (3) even when interest rates varied within institutions, the differences were almost always tied to different products serving different kinds of customers; again variation was hard to exploit. (For their part, market surveys always indicate that borrowers want cheaper credit, but it's not clear how strong borrowers' sensitivities are.)

Dehejia, Morduch, and Montgomery (2012) made a first attempt to estimate the elasticity of loan demand in a non-randomized difference-in-difference framework, exploiting a quasi-experiment (not an RCT). *Safe*Save, a lender in the slums of Dhaka, had charged its customers 2 percent per month for loans, but they felt that rates had to be increased to 3 percent to cover costs.[8] So when new branches opened, *Safe*Save charged 3 percent there. Eventually, the older branches were brought into conformity with the new branches, giving a chance to see changes in loan demand as interest rates were increased from 2 percent to 3 percent per month in the older branches. Loan demand in newer branches could then be used to control for macro shocks and broader conditions in a difference-in-difference framework. The situation was unusual in that prices were raised in some branches but not others, keeping all else the same.

Counter to the assumption that the elasticity would be zero, Dehejia, Morduch, and Montgomery (2012) estimate a long-term elasticity over -1.0. In other words, raising the interest rate by 10 percent led to a greater than 10 percent drop in demand. Rather than greatly expanding revenues, the interest rate hike slightly

[8] Disclosure: At the time, I was a member of the *Safe*Save cooperative, effectively serving as a member of its governing board. The institution is now part of the NGO BRAC.

undercut net revenues and reduced borrowing. The study directly contradicted expectations—and an important pillar of microfinance thinking. In the first rigorous test, customers were shown to care about interest rates, and they borrowed less as prices rose.

The study rested on strong assumptions. Most important was the assumption that the timing of the move from 2 percent to 3 percent was effectively random: that it was independent of demand patterns in the branch. The case relied solely on the recollection of the lender's chairperson. A case also had to be made for comparability across branches in order to interpret the difference-in-difference, relying on a demonstration of the similarity of pre-change trends. In addition, the result was based on evidence on the choices of 5147 members of a particular institution in just one set of branches in the Dhaka slums, and it was not clearly generalizable.

Still, the result mattered because it was plausible and so sharply countered expectations of practitioners (if not of economists, who take it on faith that most often demand curves slope downward). The study laid out an argument that poor households, particularly the poorest, *did* take price increases into account—and reduced loan demand accordingly.

A broader case was provided by an RCT. Karlan and Zinman (2019) describe a similarly motivated study of Banco Compartamos in Mexico. The bank is the largest lender in Latin America, serving millions of borrowers, rather than the thousands served by *Safe*Save. Compartamos is known as one of the most commercially focused micro-lenders, charging interest rates around 100 percent per year (Rosenberg 2009). The bank wanted to reduce interest rates, and Karlan and Zinman (2019) saw the chance to estimate interest rate elasticities by convincing Compartamos to reduce interest rates to different levels in different places, creating randomized treatment and control groups in the process (just as the *Safe*Save study also needed heterogeneity across branches).

Randomization at Compartamos proceeded at the branch level, covering branches spread across Mexico. Forty regions were randomly assigned to a "high rate" group: their loans cost about 10 percentage points below existing interest rates. Another 40 regions were randomly assigned to a "low rate" group with loans costing about 20 percentage points below existing interest rates. The study assessed elasticities by comparing loan demand across branches.

As with Dehejia, Morduch, and Montgomery (2012), borrowers were shown to be sensitive to interest rates. Karlan and Zinman (2019) estimated an interest rate elasticity after the first year of −1.1. By Year 3 it was −2.9. Moreover, as with Dehejia, Morduch, and Montgomery (2012), the move did not obviously help profits. After the price change, Compartamos had more borrowers but more costs too.

Had the researchers not intervened, Compartamos would likely have reduced interest rates everywhere at the same time, leaving no control group. And if Compartamos had instead deliberately chosen some branches as first-movers, the

risk of selection bias would have arisen. The RCT thus created analytically useful variation. The use of randomization by Karlan and Zinman eliminated the challenge in comparing behavior across branches. It also eliminated the concern that the choice to reduce interest rates (and by how much) in any particular branch was driven by local conditions. Rather than yielding plausible estimates that rely on a chain of assumptions (as in Dehejia, Morduch, and Montgomery 2012), the RCT parameters estimated by Karlan and Zinman (2019) are transparent and tightly measured.

On the other hand, Compartamos is unusual: the baseline interest rate was very high, and the policy change involved reducing interest rates rather than increasing interest rates. As with the result from *Safe*Save, the estimates are not directly exportable to other settings. Nevertheless, the two studies together can shift priors in a Bayesian sense, and, the experimental mindset behind the Compartamos RCT allows us to see something that would have otherwise been hard to see.[9]

3.5.3 Poverty, Migration, and Mobile Money

Technology is transforming the financial landscape, taking focus away from traditional microcredit, but use of technologies like mobile money (using telephones to make payments and maintain digital wallets) is highly self-selected. The choice to adopt new technology is reinforced by policies by providers to focus on the most lucrative parts of markets. Most often that means that poor households are disproportionately excluded. The corollary is that the poor households that adopt tend to be unusual. How then to assess the possibilities for technology in poor communities?

The rural population of Bangladesh has been steadily drawn to Dhaka, largely driven by the hope of employment in the ready-made garment industry. Factories, large and small, are following China's lead and exporting globally. The jobs are often filled by younger workers who support families in the countryside. This dynamic is in the spirit of the Lewis (1954) model of rural–urban migration and economic growth, and Bangladesh has been growing at about 6–7 percent per year. But where does this leave households remaining in rural areas? One question is whether technology can help migrant workers in Dhaka send money back to their families? Can the technology lead to increases in levels of remittances

[9] The RCT doesn't answer all questions. It appears that much of the increase in lending was due to new borrowing (not substitution from other sources), but there remain questions about impacts on well-being and risks of over-indebtedness. Moreover, the RCT is limited in what it can reveal about context and heterogeneity. The results also say nothing about the ethical questions surrounding charging relatively high interest rates to poor borrowers (Rosenberg 2009).

from urban migrants to rural families? Can it be a mechanism to reduce poverty and spatial inequality?

Lee et al. (2020) use an RCT based on an encouragement design to study how access to mobile banking changes lives in very poor communities in Bangladesh. We started with a sample of households in the rural northwest that were determined to be "ultra-poor," a group that suffers especially during the *monga* (lean) season. The households had participated in a program with a local NGO that helped their adult children move to Dhaka factories. The study follows both sides of the remittance equation, senders and receivers. In Dhaka, we followed urban migrants originally from the northwest. In the northwest, we followed their extended families. In the control group, just 11 percent had bank accounts and 20 percent were actively using mobile money.

One reason for low initial adoption of the technology was the hurdle created by English-language menus on the telephone interface used by the mobile money providers. The main experimental intervention, designed by the researchers, involved training randomly assigned groups in both urban and rural settings about how to use the technology. Participants were given hands-on experience with sending remittances, received translated menus, and got assistance with account sign-ups. (The training cost about $12 per household.) The control group received neither training nor help.

The first result was the finding of a large increase in active mobile money usage, from about 20 percent in the control group to 70 percent in the treatment group. Remittances from urban migrants back to their rural families increased by 30 percent relative to the control group. That flow of money led to a drop in extreme poverty in the rural area. Average consumption increased by 7 percent on average relative to controls, and gains were particularly notable during the lean season. Migrants, on the other hand, were more likely to report diminished physical and emotional health, consistent with pressures to work longer hours and increase remittances enabled by the mobile banking technology.

The experiment behind the RCT reduced barriers to entry for particularly excluded groups. That might have happened eventually without an RCT, but the experimental intervention allowed a clear comparison to a similar control group at a historical moment when causal inference was possible. By centering on the migration–remittance relationship, the study presents an alternative path to improving rural conditions. Standard responses are to bring resources into rural areas through microcredit and "graduation" programs which aim to raise productivity in rural areas (see the RCT by Bandiera et al. 2017). Here, instead, the mechanism involves helping rural workers find more remunerative employment in cities—and then facilitating a mechanism to move resources from the city to rural areas.

While this might seem to be a "what works" evaluation, the study is better seen as an inquiry into spatial inequality and whether intra-household sharing is

limited by costs. The point of the study is not to show that illiterate Bangladeshis are deterred by English-language menus required for operating mobile money accounts. That is not a surprise, and would not have been worth studying so intensively. Instead, the point was to use that hurdle (and a training program to overcome it) as a way to induce variation in who uses mobile money and who does not. In other words, the hurdle was the key to forming a treatment and control group (through an "encouragement design") that allowed the mapping of the consequences of access to mobile money for migrants and their families. In the end, the study does not promote a particular solution so much as contribute to understanding the channels of exit from rural poverty.

3.6 Market Failure and Private Goods

RCTs by nature are particularly useful in studying discrete interventions. They are particularly well-suited for inquiries around the delivery of private goods. The examples above are in that line. By the same token, RCTs are far weaker in assessing the role of public goods and macro change (Hammer 2017).

Some criticize RCTs for pushing the focus of development economics toward the provision of private goods, but this orientation within development economics and development policy emerged decades before RCTs came to the fore. The 1970s saw a fundamental shift in development economics toward concern with the provision of private goods. This came in the context of a broader shift toward concern with rural development, absolute levels of poverty, undernutrition, high mortality rates, and low educational attainments. The shift can be seen in the "basic needs" literature and criticisms of growth-based development (e.g., Chenery et al. 1979), the re-orientation of the World Bank under Robert McNamara, the rise of information economics within development economics (Stiglitz and Weiss 1981, Stiglitz 1986, Bardhan 1984), and a focus on "merit goods" (Musgrave 2008).[10] United Nations Millennium Development Goals and the Sustainable Development Goals—with their focus on poverty, health, education, and basic rights—reinforce the focus. One reason that RCTs took hold is because they are particularly well-matched for inquiries about the delivery of key goods and services.

As Rodrik (2009) and Ravallion (2012), note, this puts the focus on fairly small interventions, not on the larger macro changes that drive poverty, inequality, and economic growth. Restricting attention to interventions that can be studied with RCTs, critics argue, impedes attempts to bring systemic reform in places where systems are badly broken, distorted, and unfair. To put it too sharply, RCTs are

[10] Schooling is included here as a private good because, unlike typical public goods, schooling is largely "rival" and "excludable." Since there are clear externalities for the larger community, schooling is perhaps best thought of as a merit good.

particularly good for studying the impact of band-aids, and as a result we will have many studies of band-aids. RCTs are also particularly good at investigating delivery mechanisms ("last mile problems") rather than large, sectoral policy priorities (first mile problems?). Instead, critics argue, we need to tackle the structural inequalities, environmental conditions, political imbalances, and weak infrastructure that generate and reproduce the harms that band-aids can only cover up.

The critics make a fundamentally important point, and perhaps it is well to stop there. But stopping there risks ignoring a broader history, a deeper conflict, and important, unanswered questions about the roles of band-aids and delivery mechanisms, knowledge, and progress.

First, this framing makes explicit that what often takes the form of technical, statistical debates about the appropriate methods to ensure internal validity and external validity is instead most fruitfully recognized as part of a political debate about the scope and nature of intervention. The technical debates can be resolved on their own terms—and are being resolved on their own terms through statistical innovation and improved research designs—but that cannot resolve the more fundamental political tensions about the scope of intervention.

Second, the theoretical argument for systemic reform is compelling. The massive reductions in global poverty in recent decades, for example, have resulted from broad, systemic change, especially in Asia (Ravallion 2012). Yet, systemic change is not always possible, and sometimes leaves parts of populations behind. Broadening access and service delivery, and expanding the provision of basic goods, remains a fundamental agenda for governments, aid agencies, and foundations.

One might reasonably argue that development economics *should* be much more focused on context and on public goods (Hammer 2017), macro interventions, and other kinds of policy, but it is misleading to argue that RCTs are at the root of perceived imbalances. The political economy and history run deeper, and there continue to be justifiable reasons to focus on improving the delivery of private goods and services (even absent RCTs). The RCT results will not spur revolutions, but they can, cumulatively, create necessary steps to better outcomes.

3.7 Conclusion

Debates on RCTs are often unsatisfactory. They fail to distinguish between types of RCTs and types of questions. Much of the criticism of RCTs is compelling both on philosophical and technical grounds, and critics rightly argue that RCTs are not a uniquely valuable source of credible impact evaluations. Other methods are useful too, and sometimes superior. We need more description, more qualitative data, more big data, more studies with other empirical strategies.

At the same time, however, the terms of debate fail to emphasize what is truly innovative and exciting about RCTs. First, all imperfect approaches are not

equally imperfect. Adding new tools like RCTs broadens the scope of methodological possibilities. Second, often the setting needs to be shaken up in order to see something.

Randomistas emphasize the role of RCTs in determining what works and what does not. I have instead focused on those RCTs that pull economic structures apart. The difference between the two kinds of RCTs above—RCTs for evaluation versus RCTs for exploration—is the difference between studying what exists versus tinkering and rethinking to create different possibilities to study, to push further in exposing theory to reality. Coupled with an experimental mindset, these RCTs create exogenous variation that gives a new way of seeing how important markets, institutions, and processes work.

Acknowledgement

I'm grateful for comments from Isabelle Guérin, Florent Bédécarrats, François Roubaud, Tim Ogden, and Martin Ravallion and for discussions with Tim Ogden, Michael Kremer, Lant Pritchett, and participants at the conference on RCTs in Development at Agence Française de Développement, March 19 and 20, 2019. Views and errors are mine alone.

4

RCTs in Development Economics, Their Critics and Their Evolution

Timothy Ogden

4.1 Introduction

Pascaline Dupas simply wanted to help some of the poorest people in the world (this story is drawn from Ogden 2017). But she discovered, as unfortunately many do, and even more unfortunately many don't, that the educational structures for training wealthy students in wealthy countries impart very few practical skills for living among or "helping" poor households in developing countries. Put another way, she couldn't get a job at an international NGO. Discouraged, she pursued her second best option: a fellowship at Harvard.

There she discovered an opportunity to move to Kenya to serve as a research assistant for a randomized field experiment. She abandoned the fellowship early to get literally closer to her original goal, helping poor people in developing countries. While living in Kenya she befriended a young mother—and watched as that woman struggled to afford needed medicine to treat her infant infected with malaria. Again, like many others, she began wrestling with the question of why. Why was it so hard for this woman to get medicine, to set aside a few dollars, to prevent her baby from contracting malaria in the first place? When friends back in France asked her what they could do to help, where they could give money to make a difference, she suggested buying bednets. When it became clear that there were few, if any, charities where it was possible to give toward buying bednets, she and a few friends (who also went on to become professional economists) set up a charity to do exactly that.

When that charity came under criticism for giving away bednets—criticism based on standard economic theories about people devaluing things that were free and free goods distorting markets—she didn't accept or reject the criticism. She decided the best thing to do was to test whether the critics were right. So she set up a randomized trial to vary the price and the implementation of bednet subsidies, to discover the optimal way to distribute bednets. The experiment was very influential by quantitative and qualitative measures. The paper has been cited more than 570 times according to Google Scholar and is in the top 5 percent

Timothy Ogden, *RCTs in Development Economics, Their Critics and Their Evolution* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0006

of all research articles according to Altmetric. It has heavily influenced policy recommendations on distribution of bednets. GiveWell, relying on this study particularly in terms of implementation of programs to distribute bednets, has channeled $100 million to free bednet distribution.[1] Overall, bednet distribution is estimated to have prevented 450 million cases of malaria and 4 million deaths (Bhatt et al. 2015; Glennerster 2016). While in the 2000s there was a great deal of debate about the effects of subsidies and free distribution of health goods—debate that had raged on for years based on competing theory and claims from non-experimental studies—today such debates have largely disappeared.

This story encapsulates most of the debate about RCTs in development economics: motivations, impact, internal and external validity, scope, theories of change, ethics, and hypocrisy. The so-called randomistas are famous for critiquing the use of anecdotes, for instance, and yet this is an anecdote. The randomistas demand evidence for causal claims, and yet there is certainly no randomized evidence to show the research in question caused anyone to change their mind or practices. Critics argue that there is "nothing special" about RCTs and yet evidence like this seems to have been more convincing to non-economists than claims based on theory or research produced through other methods.[2] Critics claim that RCTs limit the questions that can be asked and answered to narrow unimportant ones and yet saving 4 million lives is "bigger" than many economists can claim as impact from their studies of "big" questions. Critics denounce the purposelessness of "external" evaluations that don't influence policy or practice and yet this RCT grew directly out of a practice question of one of the leaders of a charity.

The debate over RCTs is as old as RCTs. This volume is at minimum the fourth—after Cohen and Easterly (2010), Teele (2014), and Ogden (2017)—to bring together conflicting views on the role of randomized field experiments in social science/development economics/policy. A review of the history of the debates would lend evidence to the dictum that science advances one funeral at a time, since no one changes their minds. I approach the topic with trepidation. My (most likely forlorn) hope for a contribution is to attempt to summarize and systematize the most prominent critiques, examine the difficulties in making a meaningful response to those critiques, and finally discuss how the practice of RCTs has evolved. Whether or not this evolution has been in response to critics is impossible to say, but the evolution *has* been responsive to the critiques. I conclude with a model for thinking about the evolution of RCTs and empirical and experimental methods, their current status, and likely future.

[1] Note: I am the Chairman of GiveWell; this data is sourced from interviews with GiveWell staff.

[2] In Ogden 2017 (pp. 201–4), Frank DeGiovanni states that the Ford Foundation funded RCTs of "Targeting the Ultra Poor" programs because the Foundation found that RCT evidence was more convincing to policy-makers (another anecdote!)

4.2 The Critiques of RCTs

Here I want to provide a brief overview of what I perceive as the main critiques of the use of RCTs in economics (and social science) to provide some semblance of order to responses. I group the critiques into seven categories:

1. The "Nothing Magic" critiques
2. The Black Box critiques
3. The External Validity critiques
4. The Trivial Significance critiques
5. The Policy Sausage critiques
6. The Ethical critiques
7. The "Too Much" critiques

There are other critiques and nuances of these critiques that I do not cover. I'm certain that some critics will object to the way I characterize their arguments. But one must start somewhere.

4.2.1 The Nothing Magic Critiques

This critique is so named because its most direct expression is "There is nothing magic about RCTs." This critique is a response to the idea that RCTs "sit atop a hierarchy of methods" (Ravallion, Chapter 1, this volume) for estimating causal impact.

The main version of the Nothing Magic critique is that randomization does not necessarily yield a less biased estimate of impact than other methods. Deaton and Cartwright (2018) is the most complete discussion of the main form of the Nothing Magic critique. Wood (2018) details 26 assumptions required to believe that an RCT in fact yields an unbiased estimate. This critique often points back to or builds off of the debates around RCTs extending back to Student's (1938) critique of Fisher.

Another version of the Nothing Magic critique is that field experiments in economics do not conform to the double-blind standard of RCTs in medical practice—and could therefore be referred to as there is "nothing magic about development economics RCTs." The inability to run double-blind trials, or even blind trials, means that RCTs in social sciences generally don't meet the requirements to reduce one of the main sources of expected bias.

A third version of the critique says that even if RCTs do limit degrees of freedom, nothing is eliminated. Therefore RCTs have to be as carefully scrutinized as other methods. Ioannidis (2018) summarizes the results of several reviews of

RCTs that find significant evidence of bias in published RCTs in several disciplines. Young (2019) finds that the majority of published RCTs fail to correct for multiple testing and fail retroactive tests for significance. Kaplan and Irvin (2015) study the results of medical trials using RCTs and find that the number of "no-effect" results increased markedly when researchers had to file a pre-analysis plan documenting exactly how they would assess the data gathered before the experiment was conducted. Furthermore RCTs are as vulnerable to false positives, false negatives, and magnitude errors as any research method (Gelman 2018).

4.2.2 The Black Box Critiques

Closely related to the Nothing Magic critiques are a set of critiques positing what can be learned from most RCTs is limited to whether some intervention "worked" but not *why* it worked. An RCT does not necessarily illuminate the actual casual mechanism even when a causal relationship is convincingly established. A useful example is an interview of James J. Choi and Dean Karlan about an RCT they conducted where the two (and the implementer) disagree about the root cause of their results (Dubner 2018). RCTs that find no effect can be even worse. It many cases it is impossible to determine whether the null result is because of an ineffective treatment or an ineffective implementation of the treatment. Interpretation of null results is often unclear even among proponents of RCTs (Evans 2016).

A variation of the Black Box critique is the "theory-less" critique. An RCT that is not grounded in theory can be very difficult to interpret regardless of whether the outcome is distinguishable from zero or not. But if there is a well-grounded theory informing the RCT, the benefits of randomization may be quite limited. It is possible to conceive of alternative (and simpler to implement) approaches to test a clear theory.

4.2.3 The External Validity Critiques

The External Validity critiques point out that each RCT is anchored in a highly specific context. This includes such things as the implementer carrying out an intervention, often an NGO, the personnel hired by that NGO, local and regional culture and customs, the survey technique, the specific way questions are asked, even the weather. Thus, while the results from a particular RCT may tell you a lot about the impact of a particular program in a particular place during a particular point in time, it doesn't tell you much about the result of even running an exactly

identical program carried out in a different context and time. An in-depth treatment of the External Validity critique can be found in Nancy Cartwright and Jeremy Hardie's book *Evidence-Based Policy*. Cartwright and Deaton also contribute to this critique in their papers (e.g. 2018), as do Pritchett (Chapter 2), and Ravallion (Chapter 1) in this volume.

4.2.4 The Trivial Significance Critiques

I term this the Trivial Significance critique to differentiate it from the common use of the term "significance" in statistical discussions. Here I use it as a synonym of "material" in business and accounting vocabulary. The Trivial Significance critique is not about statistics or relative effect size but about (at times, truly) absolute effect size: whether the programs and policies the RCT movement is focused on matter.

The critiques can take several different guises, but all share the basic point that the programs and projects measured and measurable by RCTs yield changes, even when "successful," that are not big enough to make a difference between poverty and prosperity, at anything approaching the scale of the problem of global poverty. These critiques can come from a macro perspective (the things that "really matter" are macroeconomic-level choices like trade policy which cannot be randomized) (Ravallion, Chapter 1, this volume), a systems perspective (an RCT on increasing vaccination rates doesn't improve the health system, and may in fact hinder system development) (Garchitorena et al., Chapter 5, this volume), or a political economy perspective (RCTs cannot answer whether investing in transport infrastructure, health systems, or education systems is most likely to lead to growth) (Hammer 2014).

4.2.5 The Policy Sausage Critiques

The Policy Sausage critiques are primarily associated with Pritchett. The simplified version is that policies (whether policies of government or of NGOs) are created through complex and opaque actions influenced by politics, capability, capacity, resource constraints, history, and many other factors. Policy-making is like sausage-making. Impact evaluation, and independent academic research in general, plays only a small role in the policy sausage, especially if it is impact evaluation that comes from outside the organization. Thus, the effort put into an RCT is likely wasted, as it will fail to have an effect on this complex process. Bédécarrats, Guérin, and Roubaud (2019) note the very limited number of programs evaluated via an RCT that seem to have been scaled up.

Pritchett and others argue that the process of policy change or organizational change is completely separate from the process of knowledge creation. The bridge between the two is not built on policy briefs but on painstaking work inside bureaucracies, political machines, and organizations. Pritchett has specifically claimed that the randomistas model of policy adoption is "unbelievably Cro-Magnon." (Ogden 2017).

4.2.6 The Ethical Critiques

As long as RCTs have been conducted there have been critiques that the method is unethical. There are two main forms of this critique. One is that experimenting on human beings, particularly and especially on people in poor communities as is necessarily the case in development economics RCTs, is inherently unethical. Meyer et al. (2019) find this moral intuition that experimentation is wrong is widespread in the American population, at least.

Historically, the medical research community has dealt with moral aversion to experimentation and withholding of treatment via the concept of equipoise—only conducting experiments where there is reasonable uncertainty over the benefits (or relative benefits) of a treatment (Freedman 1987). In the context of development economics, Abramowicz and Szafarz (Chapter 10, this volume) argue that the concept of equipoise has been glossed over too quickly given the poverty and deprivation of many participants in RCT studies, and is frequently simply ignored by proponents of the method.

4.2.7 The "Too Much" Critiques

The "Too Much" critique that even if there were advantages to RCTs over alternatives, those advantages do not justify the time, monetary, opportunity or "talent" costs they impose. Ravallion (2009a and Chapter 1, this volume), for instance, argues that there are too many RCTs being conducted, that they are pushing out evaluations of programs that are not suited to randomization, and implies that RCTs occupy too much space in journals. Pritchett (Ogden 2017) worries that the "main founders of the movement are all geniuses" who should be leaving RCTs to "public health PhD students at Kansas State." Deaton suggests that they are better suited for consultants to governments to settle political disputes than inform academic knowledge (Ogden 2017). Others decry that the time it takes to conduct and analyze an RCT are too high compared to other methods, even if they are imperfect. Alternatively the method generally requires organizations implementing a program to hold the intervention constant during the period of the evaluation, regardless of feedback, which imposes large opportunity costs (Whittle 2011).

4.3 The Challenge of Responding to The Critiques

The proponents of RCTs have hardly been silent in the face of these critiques. As noted there are several volumes and many more standalone papers, blog posts, interviews, briefs, books and more that present the case for RCTs and responses to particular critiques or critics. But as time has passed, the responses have been less frequent. In recent years, if anything, the proponents of RCTs have seemingly begun to simply decline to engage with the critics (though, as I will argue later in the chapter, they have engaged with the critiques, if indirectly).

Perhaps one reason is that the defenders are at something of a rhetorical disadvantage to the critics—and the situation recapitulates some of the early exchanges in the debates when it was the proponents of RCTs who were the critics of establishment methods. In brief, the critic needs simply find a few examples of a particular problem, while the response must defend an amorphous and evolving establishment.

4.3.1 What Is a Randomista?

A specific example of this difficulty is that of the basic question of defining who (or what) is a randomista. Many of the critiques are founded on what the randomistas believe—but there is certainly no manifesto or statement of beliefs or core principles that defines who is in "the club." Lant Pritchett has described the RCT movement as religious, and the emotions that are stirred up certainly seem to make that a valid comparison. To borrow Stackhouse's definition, a religion or movement is "a comprehensive worldview or 'metaphysical moral vision' that is accepted as binding because it is held to be in itself basically true and just even if all dimensions of it cannot be either fully confirmed or refuted." At first glance, particularly for anyone who has been in the audience for many public debates or discussions of the value of RCTs, this description may seem apt for both sides of the debate. But in depth discussions with the individuals quickly erodes the sense that there is a shared "comprehensive worldview" where it comes to the value, benefits, and applicability of RCTs. The interviews I conducted for *Experimental Conversations* (Ogden 2017), with 10 of the leading practitioners of RCTs, provide ample evidence of the heterogeneity of beliefs among just this small sample of RCT proponents.

Absent a statement of the core beliefs of a randomista, critics tend to rely on a rhetorical construction that can be rendered, "Since Individual X said Y at t1, group A is wrong at t2." A particular favorite of the critics is the evergreen citation of Banerjee's 2006 statement: "Randomized trials…are the simplest and best way of assessing the impact of a program." To their credit, Deaton and Cartwright (2018) cite a number of statements in addition to the standard, and particularly point to

statements from J-PAL's materials, which are a much better starting place for a critique, but are still limited unless one is relying on collective guilt or guilt by association. The point is not that such statements are meaningless or that no one believes the statements in particular, but that it is very difficult to identify who exactly would sign on to a specific statement and whether those who would or wouldn't could reasonably be defined as inside or outside the circle of randomistas.

The nearest attempt to define a randomista appears to be Ravallion (Chapter 1, this volume), who narrows in to define a randomista as someone who holds a belief that "RCTs sit atop a hierarchy of methods, who believe that RCTs are 'gold standard' for impact evaluation—the most 'scientific' or 'rigorous' approach, promising to deliver largely atheoretical and assumption-free, yet reliable, I[mpact] E[valuation]." But even this statement leaves much to be desired in terms of the precision necessary to define a group. While there are those who like Imbens (2018) (though note that Imbens is primarily an econometrician and not a development economist) would sign on to the idea that there is a hierarchy with RCTs "at the top," is it sufficient to believe there is a hierarchy or does there need to be some specified amount of space between RCTs and other methods? How would that space be measured? Even the standard economics construction of "ceteris paribus, RCTs are a superior method for impact evaluation" leaves chasms of undefined terms where there would be significant heterogeneity between "randomistas" and quite possibly less space between a purported randomista and a critic. For instance, consider the following quotations (occasionally slightly modified to remove obvious "tells"—see Appendix to Chapter 4 (this chapter) for the original unaltered quote if you would like to judge if the alterations were fair) and attempt to classify the statement as belonging to a randomista or to a critic:

1. What methods are best to use and in what combinations depends on the exact question at stake, the kind of background assumptions that can be acceptably employed, and what the costs are of different kinds of mistakes.
2. We should neither be encouraging or discouraging any particular tool just for the sake of the tool. We should be encouraging students to look for an interesting question and use the right tool to answer it. Period.
3. [V]ery good descriptive data that focuses people's attention on something they haven't focused on before has changed people's minds in policy as much as any experiment.
4. The novelty…drove the overselling of RCTs, like these silly statements that everything ought to be evaluated randomly, or the people who say they don't believe any observational evidence.
5. [No approach] makes all the problems disappear, and neither does an RCT. I don't think anybody [should think] that RCTs are magical.
6. [T]hat's part of how the randomized evaluation movement was sold to policy makers: "You're going to get answers." I don't think that's what we're

going to get. My sense is that we're going to see evaluations that are all over the map. [I]f I had to choose, I…say we should pour more energy into the big stuff than the small stuff.

7. Organizations should be able to draw on different areas to answer the relevant questions…I see a lot of crossover between different forms of causal identification…I don't think you…focus on randomized evaluations. I don't think that makes sense.
8. An impact evaluation should help determine why something works, not merely *whether* it works. [RCTs] should not be undertaken [when they] provide no generalizable knowledge on the "why" question.
9. There are many banal and useless examples of studies using every specific method.

While any one of those quotations could be set aside or disputed in some way, the fact remains that many of the purported randomistas repeatedly express sentiments that RCTs are a "tool in the toolbox" of modern economics, that there are many other useful tools, that RCTs are not appropriate for every worthy question and that other analytical tools are useful and credible. Moreover, most if not all of the randomistas use and publish other methodologies. As McKenzie (2019) has pointed out, the most cited papers of arguably the three most well-known randomistas—Banerjee, Kremer, and Duflo—are not RCTs.

The second half of Ravallion's definition ("the most 'scientific' or 'rigorous' approach, promising to deliver largely atheoretical and assumption-free, yet reliable, I[mpact] E[valuation].") would be even more difficult to get meaningful numbers of economists who practice RCTs to sign on to. There is a decade-long debate within economics of the proper ordering of data analysis and theory that is as intractable as the debate over RCTs (Cherrier 2019). "Largely atheoretical and assumption-free" could (and has) inspired pages of debate on its own. It was this very fact that led to *Experimental Conversations*, where I decided the only meaningful thing to do was to interview many of the randomistas and near-randomistas to hear them explore the nuances of their beliefs in terms of their actual practice in conducting and interpreting research, not just on the metaphysics of methodologies.

The intent of this discussion is not to end with the conclusion that the term randomista is so ill-defined and undefineable as to be practically useless, though I think that's true, but to illustrate why it is so hard to respond meaningfully to the critiques (and perhaps to explain why the randomistas have generally stopped responding directly to critics, as evidenced by the fact that none agreed to participate in this volume). Any response must necessarily be on behalf of an individual who likely will agree with at least some parts of any particular critique—and ultimately speaks only for themselves. At the same time, any person who, like I do in this chapter, tries to defend RCTs must bear some

burden to answer for any statement on behalf of RCTs no matter how off-the-cuff, uncareful, misguided or wrong. It's not surprising then that few relish, at this late stage of the ongoing discussions, the opportunity to respond to a critique of ill-defined randomistas in general with a specific statement of personal beliefs. What would that accomplish?

4.3.2 The Argument behind the Arguments

Since I am not a randomista in the sense that I do not make a career out of running RCTs, I can hardly arrogate the authority to answer the critiques or critics on their behalf. That being said, in the next section I will use actions (both studies and other professional activities) of various nominal randomistas to illustrate how the movement has evolved in ways that at least blunt many of the critiques. Before doing so, I want to point out one other issue that makes productive direct responses to the critiques difficult.

The disputes between randomistas and their discontents can put too much emphasis on the particularities of methodology, and distract from the more important disagreement behind them. That more important disagreement is about theories of change. Argument over theories of change—ideas about how the world changes—are hardly unique to the present moment in development economics. Indeed, it is the foundation of development economics (and much of other social sciences): how is it that poor countries become richer (or, why is that poor countries stay poor)?

There has always been wide disagreement within the economics profession about theories of change. One can think of most of the seminal texts in economics as manifestos about theories of change. Development economics has its own specific theory of change conflicts (e.g. Sachs vs. Easterly; Acemoglu's and Robinson's "institutions matter"; Rodrik's industrial policy; Deaton's anti-aid arguments).

While wary of reducing theories of change to short summaries or points on a chart, nevertheless I find it helpful in the context of this discussion, to think about the competing theories of change along three main axes:

> the value of small versus big changes;
> the value of local knowledge versus technocratic expertise;
> the role of individual versus collective actions via institutions.

The axes are not completely independent of each other. Someone who believes strongly in the value of big changes is obviously also very likely to place more value on technocratic expertise and the role of institutions. In practice, there is significant variation within the RCT movement and between the critics, such that

in some cases there is more in common between a particular RCT advocate and a particular critic than there is between two different critics—again the problem of definition of a randomista arises.

Underneath each of the critiques of RCTs noted above is a theory of change that differs from that of RCT advocates along at least one of the three axes. After the many conversations I've had with both randomistas and critics, my impression is that those in the RCT movement tend to believe that small changes can matter a great deal, that technocratic expertise is highly valuable, and that individuals within institutions matter as much as the institutions themselves. Those critics who invoke the Trivial Significance critique, in contrast, usually agree on the value of technocratic expertise, but disagree about the value of small changes and the role of institutions.

This difference matters because it influences how one evaluates the quality and especially the utility of evidence. If you believe that the path to improving the world is through the accumulation of small changes then external validity is a much lower bar. You do not need to be convinced that a program will yield the exact impact or even close to the same impact in a different context to be worth transferring. You simply need to believe that it provides the starting place to run another small experiment in the different locale and adjust as you go. The distinction between a randomista with this theory of change and a "feedbackista" like Dennis Whittle is simply about the speed of iteration and how much to value different kinds of feedback. Implicit there, of course, is that this same randomista would look at an RCT that a critic calls atheoretical and be very confused—there is a theory, though perhaps not a structural model that allows informed estimation of cross-context impact. At the same time, a randomista with a slightly different theory of change that puts more value on institutions will decline to defend small-scale RCTs with NGOs and advocate for much more experimentation at scale (e.g. Niehaus 2019).

Pritchett nods to the importance of the lack of common theories of change in some of his writing and speaking: "What I worry about development is that there are two ontologically different categories to which the word is commonly applied" (Pritchett 2010a); "You can call that whatever you want just don't call it development" (Ogden 2017). But in general the lack of a shared ontological universe is not given the attention it deserves in such debates. Perhaps that's because there is unlikely to be much value in such a debate—the discussants are using the same words to mean different things.

4.4 The Evolution of the "Movement"

Among the more prominent critics of the RCT movement, such as it is, has been Lant Pritchett. In 2018, Pritchett began giving a talk he titled: "The Debate is

Over. I won. They lost."[3] In this section, I'll argue that by examining the evolution of the use and practice of RCTs it is clear that many RCT critiques have in fact been acknowledged to be correct by virtue of changing practices among RCT practitioners and research centers. In that sense, his triumphalist title is correct. However, I'll also argue that the evolution of the "RCT movement" is best explained by a model that Pritchett himself (along with Matt Andrews and Michael Woolcock) introduced (Andrews, Pritchett, and Woolcock 2012) arguing that this model was the best path to sustained development impact.

4.4.1 Problem Driven Iterative Adaption

In their original paper (which spawned a number of additional papers and ultimately a book, *Building State Capacity*) Andrews, Pritchett, and Woolcock (2017) introduce the principles of problem driven iterative adaptation:

> We propose an approach, Problem-Driven Iterative Adaptation (PDIA), based on four core principles, each of which stands in sharp contrast with the standard approaches. First, PDIA focuses on solving locally nominated and defined problems in performance (as opposed to transplanting pre-conceived and packaged "best practice" solutions). Second, it seeks to create an 'authorizing environment' for decision-making that encourages 'positive deviance' and experimentation (as opposed to designing projects and programs and then requiring agents to implement them exactly as designed). Third, it embeds this experimentation in tight feedback loops that facilitate rapid experiential learning (as opposed to enduring long lag times in learning from ex post "evaluation"). Fourth, it actively engages broad sets of agents to ensure that reforms are viable, legitimate, relevant and supportable (as opposed to a narrow set of external experts promoting the "top down" diffusion of innovation).

These four principles are clearly seen in the evolution of the RCT movement.

Principle One: Solving Locally Nominated and Defined Problems in Performance

Michael Kremer is generally credited with beginning the use of RCTs in development with a randomized experiment on the impact of textbooks on learning in Kenyan primary schools; the only controversy in this regard is that Santiago Levy was nearly simultaneously deploying an RCT to evaluate the impact of Progresa, a new conditional cash transfer program in Mexico. Regardless of which of the

[3] The title slide reads "We won. They lost." but in his remarks he uses "I."

two is credited with initiating the use, both projects embody the first principle. Levy's motivation was solving a particular local problem. At the time the program was created, it was clear that the then current government of Mexico was going to lose the next national election. Levy was concerned that the program would be canceled by the incoming government. He implemented an RCT to establish the impact of the program in order to protect it from political concerns.[4]

Kremer's first RCT was motivated by a discussion with a Kenyan friend, Paul Lipeyah, who had the job of picking seven primary schools to receive new textbooks from the NGO ICS. Kremer describes his thinking like this:

> In 1994 when I started the work in Kenya, I was very much influenced by the movement for the better identification in labor economics and public finance… I also was not reacting to the critics of instrumental variables. Indeed, I think those working on instrumental variables and those of us working on RCTs were motivated by the same impulse, the concern that a lot of empirical work in economics at the time was potentially subject to confounders and required a lot of fairly strong assumptions. That being said, it's not like IV makes all the problems disappear, and neither does an RCT. I don't think anybody thinks that RCTs are magical, but they are a really useful tool for getting at causal impact. So I would say I was trying to get at causal impact in a way that was part of a broader movement in the economics profession to get better identification …My main impulse was practical—to get more believable answers to real world questions. I have always been mainly interested in the underlying questions of what policies can address poverty and I realized that RCTs were a tool that could be adapted to help answer this question. I was motivated to make RCTs a more flexible and useful tool. (Ogden 2017)

Kremer was clearly also trying to solve a locally nominated and defined problem in a double sense. First, he was trying to help the specifically locally defined problem of ICS of picking which seven schools would receive textbooks. Second he was addressing the locally defined and nominated problems among economists of improving causal identification.

Similar stories accompany the entry into RCTs of other economists who are well-known implementers of RCTs. Pascaline Dupas' story, which opened this chapter, is a good example: the first RCT she conducted was in response to the locally defined and nominated problem of whether it was better to give away bednets or charge for them. David McKenzie's first RCT, a test of the returns to capital for microenterpreneurs in Sri Lanka, came about because he had already done similar non-experimental work in Mexico, "But people were not convinced by the nonexperimental results." (Ogden 2017). He was motivated by the locally

[4] Author interview with Santiago Levy, 2018.

defined and nominated problem of convincing economists and policy-makers that microentrepreneurs could have significant returns to capital. Sometimes, as with Dupas's and Kremer's story, the locally defined and nominated problem was a question that was shared by the economist and an NGO or government agency.

Principle Two: Create an 'Authorizing Environment' for Decision-making that Encourages 'Positive Deviance' and Experimentation

It's clear that the initiators of the use of RCTs created an authorizing environment that encouraged positive deviance and experimentation, the last quite literally. The level of innovation within the conduct of RCTs is quite impressive. From generally small-scale experiments on very simple interventions, e.g. textbook distribution to seven schools, the practitioners of RCTs have innovated and evolved consistently. From a methodological standpoint, the process of both randomization and analysis has become much more sophisticated to deal with highly varied contexts and potential confounders of prospective balance, multiple hypothesis testing and other potential biases. Early implementers of RCTs such as Ted Miguel have been leaders in research transparency, data and analysis disclosure, and the use of pre-analysis plans. A "second wave" of randomistas have gone on to implement much more sophisticated experiments over much longer timeframes (e.g. Blattman and Dercon's experiments comparing industrial jobs to microcredit) and much, much larger scales (e.g. Muralidharan and Niehaus's experiments with NREGA and Aadhar) on much more complex topics (e.g. Pomeranz's experiments on taxation schemes and Karlan's experiments on religious content in an intervention).

Principle Three: It Embeds This Experimentation in Tight Feedback Loops that Facilitate Rapid Experiential Learning

The only argument about the randomistas' implementation of this principle is whether it is something they directly created or an extant feature of economics education that they exploited. I would argue it is both. The nature of economics education and practice means that each new generation of economists learns by doing the grunt work for the prior generations. It is hard to imagine a tighter feedback loop to and more rapid experiential learning than a Ph.D. candidate (or even earlier) spending a year or two as a field research manager on a variety of experiments being overseen by existing practitioners. RCT implementers have also exploited the network of events that the economics profession has as an opportunity for feedback loops on innovations in field experiment set-up, management, and analysis—conferences like the Northeastern Universities Development Consortium (NEUDC) and Pacific Development Consortium (PacDev) have, much to the chagrin of RCT critics, often become platforms for rapid learning in how to conduct, analyze, and report RCTs.

Principle Four: It Actively Engages Broad Sets of Agents to Ensure that Reforms Are Viable, Legitimate, Relevant, and Supportable

The institutions to support the implementation of RCTs are the best examples of this principle in practice. Since the first RCTs, prominent users of RCTs have created organizations like J-PAL, IPA, and CEGA which easily match the description as broad sets of agents that ensure that reforms are viable, legitimate, relevant, and supportable. All of the organization are involved in ongoing projects to reduce the barriers to the conduct of and reporting of RCTs. These include books and courses about how to conduct RCTs, training of many students within and outside of universities, and creation of research organizations in developing countries (e.g. the Busara Center for Behavioral Economics, and permanent field staff in a number of countries).

4.4.2 PDIA-driven Evolution and Critiques of RCTs

True to Andrews, Pritchett, and Woolcock's promise, the implementation of PDIA has served to vastly improve the practice of RCTs, their relevance to development practice and to policymakers, and to institutionalize the process of conducting and reporting the results of randomized experiments. But two additional points are necessary.

First, the practice of RCTs has developed as practitioners confront the problems that they perceived. The "locally nominated" problems that the evolution has confronted are primarily the problems faced by RCT practitioners—publishing research and fulfilling personal career objectives. As I will argue later, many of these career objectives are about improving the world and helping the world's poorest. But that does not obviate that the process of evolution is driven by motivations internal to the movement.

Second, is that this process of internally driven improvement is, according to Andrews, Pritchet, and Woolcock, the only reliable and sustainable way to build capacity. Pritchett often laments that many of the "locally nominated problems" that randomistas have been attempting to address were "entirely predictable" (Ogden 2017) and yet the randomistas ignored the critics. But as the quotation above introducing PDIA says, "transplanting pre-conceived and packaged 'best practice' solutions" and change based on a "narrow set of external experts promoting the 'top down' diffusion of innovation" simply doesn't work. The only way for the RCT movement to evolve into a sustainable and effective force for development was to develop capabilities and solutions internally.

In this section, I'll briefly discuss how the PDIA process has led to the evolution of the practice of RCTs to at least in part address many of the critiques of the movement.

Nothing Magic

As noted in an earlier section, it is not clear how many RCT practitioners ever believed that RCTs were magic or were not subject to any biases. It is worth noting that Brodeur et al. (2018) find significantly less evidence of p-hacking and significance searching in RCT and RDD papers than in IV and Difference-in-Difference papers and Vivalt (2019) finds less significance inflation in RCTs than in papers using other methods. Perhaps even more important in relation to the present discussion, she finds that RCTs have "exhibited less significance inflation over time."

It is however likely that many RCT practitioners did not appreciate the many sources of bias that remain in randomized experiments when they first began. If there were believers that RCTs solved all of the problems that critics point out, we would expect that RCT practitioners would resist innovations in implementation and analysis that better account for such sources of potential bias.

In fact, what we see is many economists who might be called randomistas are actively innovating to address concerns about bias, reliability, and replicability. Several are worth specific note.

> Ted Miguel is one of the founders of the Open Science Framework and of the Berkeley Initiative for Transparency in Social Sciences. Both organizations encourage researchers to make all the data and code used in their work accessible for replication (whether RCT or not).
>
> Randomistas including Dean Karlan, Esther Duflo, and Chris Blattman have been vocal advocates of pre-registration of studies, especially RCTs, and were involved in creating the AEA RCT Registry.
>
> J-PAL has started a replication service where "a graduate student attempts to replicate a complete paper from scratch, and can identify any error, omission, or questionable assumption" (Crepon et al. 2019).
>
> The World Bank's Development Impact Blog ("news, methods and insights about impact evaluation") frequently features critiques of papers using RCTs, advice on new statistical methods and techniques for improving RCT analysis, and on non-RCT methods.
>
> Guido Imbens, cited above as saying that RCTs do in fact sit on top of the hierarchy of methods, continues to work on methodological improvements on other methods, such as difference-in-differences and machine learning (e.g. Athey and Imbens 2018, 2019)

Each of these examples can be seen as examples of each of PDIA's principles, given that they address problems of performance in conducting and interpreting RCTs, and are conducted without any central authorizing authority or mandate.

More importantly, these efforts illustrate that far from treating RCTs as magic, those within the loose borders of the RCT movement recognize sources of bias and error in RCTs, and actively invest in addressing them. Similarly, they continue to employ methods other than RCTs, and do not, in practice, ignore work on other methods.

Black Box

Again there has been considerable evolution in the application of RCTs along PDIA principles to address the limitations of simple RCTs in exposing causal mechanisms. A few examples include:

> Alfonsi et al. (2017) study youth employment schemes in Uganda comparing vocational training programs to subsidies for on-the-job training. They not only establish that vocational training has a higher impact in terms of youth income, but that the mechanism is greater mobility between employers because of certifiable skills, ultimately leading to better matching of certified youth to higher productivity firms.
>
> Beaman et al. (2018a) study agricultural technology diffusion through social networks in Malawi and establish the nature of the "complex contagion" that leads to farmers adopting new methods, and use the mechanism to identify ways to cost-effectively improve targeting of agricultural extension programs.
>
> Cai and Szeidl (2017) experiment with Chinese business networks finding that networking meetings significantly improve firm performance, specifically through peer learning from better functioning firms on topics outside the intervention and in better supplier-client matching, and that the regularity of meetings matters.
>
> Campos et al. (2017) study small enterprise training programs in Togo, comparing traditional business training to personal initiative training. They not only find that that the personal initiative training is more effective in boosting firm profits, but the specific behaviors that changed including product diversification, innovation, and investment.
>
> Karing (2018) studies a program to encourage child vaccination in Sierra Leone and not only establishes the efficacy of public signaling through colored bracelets, but that the mechanism is social desirability (not attention) and that the effect of the bracelets varies based on the social desirability of specific vaccines.

The examples chosen here are not systematic but are deliberate to illustrate that an emphasis on addressing the "black box" critique is not limited to a few researchers, schools, contexts, or sectors. As more such studies are done, the implicit PDIA process operating within the RCT movement means that establishing mechanisms will increasingly be expected of new RCTs.

External Validity

The external validity critique would in general be more credible if some of its proponents were as vocal about the problems of external validity of all studies and not just of RCTs. As Pam Jakiela has noted in response to Cartwright and Deaton, "Nice insight by Deaton and Cartwright, but for some reason they keep spelling 'study' as R-C-T."[5] That being said, there are many instances of RCT proponents offering policy advice which assumes external validity.

Faced with the problem of proving external validity, RCT practitioners have evolved in several ways. First they have empirically studied whether the results of RCTs in one context predict results in another. For instance, Meager (2019), using Bayesian Hierarchical Modeling (another example of the application of PDIA to the "Nothing Magic" critique) shows that the variation between RCTs of microcredit is smaller than it appeared (Pritchett and Sandefur 2015) and therefore the results of the individual RCTs are reasonably predictive of the results in other locations. Alcott (2015) does something similar comparing the ability of an RCT of reminders to reduce energy consumption in one city to predict the effect of the same campaign in another city finding that RCTs don't do a great job, but a better one than other methods in common use.

At the same time, RCT practitioners have put much more emphasis on replications and studies with multiple arms in multiple contexts. For instance Dupas et al. (2018) test savings encouragements in Uganda, Malawi, and Chile. Perhaps most famously, a wide variety of researchers collaborated to test Targeting the Ultra Poor programs in eight locations with both government and NGO implementations. This "replication" for external validity can also take place significantly ex-post. For instance, Bernhardt et al. (2017) reanalyze data from multiple experiments on differential returns to capital between male and female entrepreneurs to identify a previously unclear causal mechanism relating to household bargaining and optimization (which is also an example of addressing the Black Box critique). Such ex-post replications are becoming easier because of the efforts of other RCT implementers to ensure data and code for all experiments are available for replication.

Of course, there will always be questions of external validity in the application of any impact evaluation (RCT or otherwise) to predict outcomes in other contexts. But more systematic approaches are also evolving. As more RCTs address causal mechanisms, assumptions about external validity will become more explicit, and more studies will include structural models. This in turn, will allow more formal frameworks for assessing external validity and integrating results from multiple studies such as Dehejia, Pop-Eleches and Samii (2019) and Wilke and Humphreys (2019).

[5] https://twitter.com/PJakiela/status/797053999925104640

Trivial Significance

Earlier I noted that a key foundation of the trivial significance critique is differing theories of change between randomistas and critics. Little can be done to respond to a critique that the only changes that matter are macro-level policies. That, however, is more a critique of applied microeconomics in general than RCTs. That being said, the RCT movement has a response to at least one of the critiques emanating from a different theory of change. For those that argue that institutions matter, the institution building prowess of the randomistas should be impressive. Aside from the obvious examples of IPA and J-PAL, the institutions built by the randomista movement directly and indirectly include the Global Innovation Fund, the Busara Center for Behavioral Research, 3ie, Evidence Action, Development Impact Ventures, AEJ: Applied, and many local survey firms.

Here though I want to focus on a variety of the trivial significance critique which is grounded in the original experiments that popularized the use of RCTs—textbooks in schools, getting teachers to show up for school, incentivizing vaccination, etc. These critiques focus on both the small nature of the intervention and the small measured results even when statistically significant (see Harrison 2011 as one example). Another related variation laments that RCTs are not well suited for measuring long-term impact (see Ravallion 2020).

Confronting these issues has yielded an impressive amount of creativity in application of RCTs. The most direct response to the "too small" critique has been expanding the scale of RCTs. Muralidharan, Niehaus, and Sukhtankar (2016, 2018) offer the best example by studying a safety net program in Andhra Pradesh, India, with 19 million people in the experiment, with an estimated savings of $38.5 million per year. While the fact that both figures are in millions rather than billions will leave some unsatisfied, it's certainly marked progress over the early years of the RCT movement.

Other randomistas have pushed the boundaries of what can be studied via RCT in other ways:

> Dina Pomeranz, with a variety of co-authors, has conducted RCTs on a variety of tax policy questions.
>
> Chris Blattman, with a variety of co-authors, has randomized access to factory jobs in Ethiopia, policing strategies in Colombia, and anti-violence campaigns in Liberia.
>
> Bryan, Choi and Karlan even conduct an RCT on the impact of religious belief, randomizing Christian evangelism in the Philippines (finding support for the Protestant work-ethic hypothesis).

Policy Sausage

The translation of RCT results into policy changes has always been an explicit goal of RCT practitioners. Their stories of how and why they began running RCTs

commonly include a pragmatic desire to influence policy in order to make a concrete difference in people's lives.[6] A quotation attributed to Michael Kremer by Karthik Muralidharan is illustrative: "Never apologize that your fundamental motivation is to improve the lives of hundreds of millions of people, and that economics is a tool to get there and not an end in itself."[7]

Their good intentions notwithstanding, the initial work of the randomistas in terms of policy influence could be described as Pritchett has on various occasions: naïve, Cro-Magnon. As Bédécarrats, Guérin, and Roubaud (2019) note, less than five percent of RCT impact evaluations conducted by J-PAL have led to scaled-up policy changes.

But as time has passed the sophistication and intensity of efforts to affect policy has accelerated. Having identified this locally nominated problem of limited policy impact, the RCT practitioners rapidly iterated in an environment that encouraged positive deviance and provided rapid feedback within the group.

The initial assumption that evidence would generate policy change mechanically has fallen away in favor of focused efforts to influence policy. This includes the creation of policy-focused teams at both J-PAL (including a "government innovation" initiative at J-PAL that works specifically to support government agencies conducting policy implementation experiments) and IPA. But it also includes participation in the creation of standalone organizations to implement programs based on RCT evidence (namely, Evidence Action), organizations to encourage the creation of and use of evidence in policymaking (3ie), internal groups within existing policy and implementation organizations (Development Impact Ventures at USAID), close collaboration with research groups at NGOs (BRAC, Pratham), education programs for policy makers and implementers, and of course, training a huge number of masters and Ph.D. students in the methods and approaches, the vast majority of which will end up in policy-related jobs rather than in academia. Some practitioners have even taken on roles in the policy-making apparatus—Rachel Glennerster's role as Chief Economist at DfID and Andrew Leigh's role as a parliamentarian in Australia come to mind.

Put another way, the randomistas have engaged a broad set of agents to ensure the validity and continuity of the use of RCTs to influence policy. A new generation of RCT practitioners are going to be an integral part of policy-making institutions (if for no other reason than the shortage of jobs in academia).

The Ethical Critiques

There is much less to say on this topic. In part, that is because fundamentally randomistas clearly believe that experimentation with human beings is ethical,

[6] In Ogden (2017), see interviews with Michael Kremer, Esther Duflo, Dean Yang, Chris Blattman, among others.

[7] https://twitter.com/karthik_econ/status/1102237584103600129

regardless of the moral intuitions of the majority of American public, an attitude of course shared by most scientists. The common refrain, which I am of course sympathetic too, is that there isn't a choice about whether to experiment (since every policy implementation is an experiment) there is only a choice of how much is learned from an experiment.

But clearly there remain many questions about the ethics of experimentation. During the summer of 2019, a new working paper that randomized encouragement to participate in anti-authoritarian protests in Hong Kong (Bursztyn et al. 2019) attracted a huge amount of attention[8] specifically because many economists seem to believe the experiment was unethical. An oft-asked question, on Twitter at least, was how the experiment managed to apparently pass through several Institutional Review Boards. The paper and the subsequent discussion revealed[9] that there are yet no meaningful bounds or codes or even shared principles on where economists should draw an ethical line in terms of experimentation.

On the questions of equipoise, as noted above, this remains an area where the RCT movement has yet to significantly engage as best I can tell.

Too Much: The Final Critique

The final category of critique I identified falls outside of the PDIA framework as it is not a critique of what RCT practitioners in development economics do, but of how much they do it. I find this the least compelling of all critiques within the economics frame.

To begin with, as many of the Too Much critiques acknowledge, the emergence of RCTs in development economics is in no small part due to the conditions and structure of the market for academic economics. The use of RCTs gained popularity in the context of widespread questions about the credibility of other methods, in an environment that demanded of aspiring economists that they do work that was credible, novel and publishable. RCTs promised—and delivered—work that was all three. Thus the criticism of Too Much should really be directed at the structures and incentives of the profession not at those who respond to the incentives the profession creates. This form of the critique is equivalent to criticizing market participants for doing the "wrong" thing, rather than addressing any market failures.

Second, the Too Much critique fails to articulate an objective measure of what the thresholds between "not enough," "just right," and "too much" might be. It is objectively true that the use of RCTs and the publication of papers using the method has increased greatly (Ravallion, Chapter 1, this volume, Bédécarrats, Guérin, and Roubaud 2019), but this growth must be put in perspective. It's worth quoting McKenzie's (2019) look at the data on this question at length:

[8] https://twitter.com/DurRobert/status/1148090885470654464
[9] https://twitter.com/arindube/status/1148807790787473410

> despite the rapid growth, the majority of development economics papers published in even the top-five journals are not RCTs… [O]ut of the 454 development papers published in these 14 [economic development field] journals in 2015, only 44 are RCTs (9.7%). The consequence is that RCT-studies are only a small share of all development research taking place.
>
> The median [BREAD affiliate] researcher had published 9 papers, and the median share of their papers which were RCTs was 13 percent. Focusing on the subset of those who have published at least one RCT, the mean (median) percent of their published papers that are RCTs is 35 percent (30 percent), and the 10–90 range is 11 to 60 percent. So young researchers who publish RCTs also do write and publish papers that are not RCTs.

Third, the oft-repeated assertion that "enthusiasm for RCTs will fade" seems to me to be a hollow critique. Of course we should expect that methods will continue to improve, new innovations in all sorts of research designs will uncover heretofore unappreciated problems and improved approaches. At some point in the not too distant future I can confidently predict that someone will write an essay about "RCTs 2.0" and make a distinction of arguable difference between the "early days" of the RCT movement and the improved methods now in vogue. Perhaps this chapter falls into that category.

Susan Athey, reacting to Judea Pearl criticizing what he terms the naïve approach to causal inference in economics (as a whole, not the RCT movement), writes: "[I] think the most effective way to evangelize a new method is to demonstrate its effectiveness in a first-rate empirical application where the method clearly leads to a better quality and more credible result. Researchers will mimic a fully worked out, successful example."[10] That could serve as a short-hand history of the use of RCTs in development economics. Enthusiasm for the original practice of RCTs has already faded as "first-rate empirical applications" of more sophisticated experiments and analysis have emerged. And enthusiasm for current practice will surely fade as "first-rate empirical applications" of improved methods—with randomization at their core or not—are created. Until then, there isn't "too much."

Finally, there is the lament that the "brightest and best" economists are wasting their talents focused on RCTs. This critique makes the least logical sense of all. If the critics are right and the problems of RCTs are insurmountable, and there are clear better alternatives, then that must indicate that those who continue to primarily use RCTs are not the brightest and best. This critique must explain why anyone should believe that the brightest and best are systematically wrong and yet still are the worthy of the moniker. And if they are not the brightest and best, why can't the actual brightest and best convince the next generation of students to

[10] https://twitter.com/Susan_Athey/status/1107422021753790464

abandon RCTs for other methods? The only plausible explanation that makes sense of this critique is that the entire profession of economics is broken, in which case the critics are wasting time on the symptoms and not the causes.

4.5 Conclusion

To conclude, I want to provide a different framework for thinking about the evolution of the practice of RCTs and the various critiques and responses. Here, once again, Pritchett and I overlap. I spent the first ten years of my career at the technology research firm Gartner. One of the organization's most widely known products is The Hype Cycle—a way of conceiving of the emergence, evolution, and adoption of emerging technologies.

The Hype Cycle posits that as a breakthrough technology emerges it passes through five distinct phases, named colorfully enough that they require little additional expectation: "Innovation Trigger," "Peak of Inflated Expectations," "Trough of Disillusionment," "Slope of Enlightenment," and "Plateau of Productivity" (see Figure 4.1).

Pritchett stumbled across the Hype Cycle and applied it to RCTs in a 2013 essay. I agree that it is a useful model for thinking about RCTs—in fact, I would

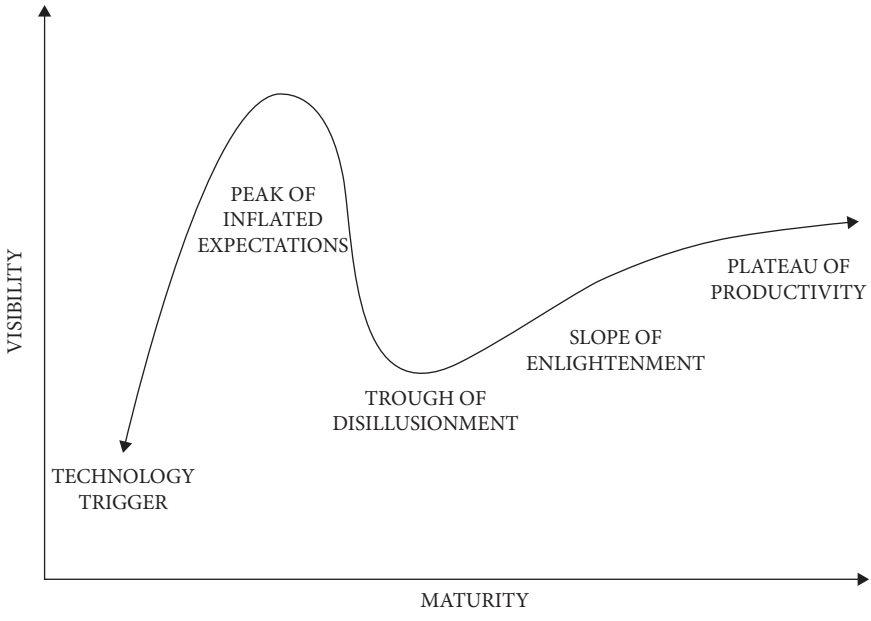

Figure 4.1 The Gartner Hype Cycle
Source: Author, based on the Garner Hype Cycle.

argue that RCTs are best thought of as an "emerging technology" in development economics rather than a movement.

Through the lens of the Hype Cycle, in this chapter I have argued that, (1) the peak of inflated expectations for RCTs was real, but was never as high as critics made it out to be, and in any event has passed, (2) initial enthusiasm for RCTs was quickly met by a range of valid critiques, leading if not to a trough of disillusionment at least to meaningful changes in the use and practice of RCTs, and (3) the current state is clearly in the slope of enlightenment phase as evidenced by the statements and practices of advanced users of the technology.

It is worth noting specifically that the evolution of the practice of RCTs validates many of the critiques detailed here. The evolution that I have attempted to document is responsive to these critiques. The practitioners of RCTs are not evolving their practice to deal with novel issues that have not been raised by critics—regardless of any direct response to the critics, the randomistas have implicitly acknowledged many of the critiques by evolving in ways that take the teeth out of many of them.

That being said, I believe there is ample reason to believe the plateau of productivity for RCTs is higher than many critics seem to make it out to be, simply as a mechanical consequence of the way the world works. There are many more decisions about implementation than there are about what to implement. Implementation decisions are clearly within the scope of RCTs. Because there are many many more students being trained in development economics than will ever hold tenured jobs at R1 universities, a significant proportion of those students will end up in jobs where implementation questions rather than larger policy questions are their purview. The training they receive in experimentation and causal identification will be highly relevant and applicable to their ability to engage in PDIA in those jobs.

Beyond that, RCTs are a more useful tool for improving the world than most tools available to the *median development economist*, given the nature and requirements of the profession, and the difficulties of policy influence. The emergence of RCT technology and the supporting mechanisms around that technology are applicable to the vast majority of the actual questions and discrete decisions about anti-poverty policies, programs, and implementation. It is true that RCTs are unlikely to be a useful tool for evaluating exchange rate policies, the optimal level of public debt, or the consequences of wealth inequality (just as a few examples). But the policies related to the answers to questions on such topics are far less susceptible to academic influence regardless of the methodology issued to answer them. As I write this (in the summer of 2019), the possibility of a massive global retreat from liberalized trading regimes is frighteningly real despite tens of thousands of macroeconomists' decades of policy effort. There is no reason to believe that the marginal impact of the average development economist studying one of these topics is greater than a precisely-estimated zero. The median development

economist's comparative advantage would be in improving the implementation of a policy or program, even without any external validity or scale-up.

In closing, I would reiterate again that Lant Pritchett was right, and he won. The critiques of the RCT movement are generally valid if not objectively correct. However, many of those critiques have been addressed by the evolution in the practice of RCTs. I expect that the evolution will continue, and that eventually RCTs may be supplanted by some other methodology (already, of course, there are new "emerging technologies" in economics: big data, machine learning and artificial intelligence—and some of the debates over the use and applications of RCTs are being recapitulated). Until then, I expect that the plateau of productivity where RCTs currently reside will continue to yield benefits to the world.

Appendix to Chapter 4: Full Quotations

1. What methods are best to use and in what combinations depends on the exact question at stake, the kind of background assumptions that can be acceptably employed, and what the costs are of different kinds of mistakes. —Angus Deaton and Nancy Cartwright

2. We should neither be encouraging or discouraging any particular tool just for the sake of the tool. We should be encouraging students to look for an interesting question and use the right tool to answer it. Period. —Dean Karlan

3. Often just very good descriptive data that focuses people's attention on something they haven't focused on before has changed people's minds in policy as much as any experiment. —David McKenzie

4. I really think despite the attractiveness of RCTs from a number of different standpoints, they're not the only methodology that we should be thinking about. There are often questions that come up that we can't answer with RCTs, and I think we can find nice credible identification strategies that are not based on RCTs but are based on other sources of data. We should be considering those questions and those studies as well. —Dean Yang

5. I think those working on instrumental variables and those of us working on RCTs were motivated by the same impulse, the concern that a lot of empirical work in economics at the time was potentially subject to confounders and required a lot of fairly strong assumptions. That being said, it's not like IV makes all the problems disappear, and neither does an RCT. I don't think anybody thinks that RCTs are magical. —Michael Kremer

6. I do think that's part of how the randomized evaluation movement was sold to policy makers: "You're going to get answers." I don't think that's what we're going to get. My sense is that we're going to see evaluations that are all over the map.

…if I had to choose, I might even say we should pour more energy into the big stuff than the small stuff.

The novelty maybe drove the overselling of RCTs, like these silly statements that everything ought to be evaluated randomly, or the people who say they don't believe any observational evidence. —Chris Blattman

7. Organizations should be able to draw on different areas to answer the relevant questions.… I see a lot of crossover between different forms of causal identification. So I think focusing, yes, but I don't think you just have to focus on randomized evaluations. I don't think that makes sense. —Rachel Glennerster

8. An impact evaluation should help determine why something works, not merely *whether* it works. Impact evaluations should not be undertaken if they will provide no generalizable knowledge on the "why" question—that is, if they are useful only to the implementing organization and only for that given implementation. This rule applies to programs with little possibility of scale, perhaps because the beneficiaries of a particular program are highly specialized or unusual,or because the program is rare and unlikely to be replicated or scaled. If evaluations have only a one-shot use, they are almost always not worth the cost. —Dean Karlan and Mary Kay Gugerty

9. There are many banal and useless examples of studies using every specific method. —Mark Rozenzweig

All quotations are from Ogden 2017 unless otherwise specified: Cartwright and Deaton (Deaton and Cartwright 2018), Gugerty and Karlan (Gugerty and Karlan 2018) and Mark Rozenzweig (McKenzie 2018)

Acknowledgement

Thanks to Jonathan Morduch, David McKenzie, Lant Pritchett, and Isabelle Guérin for helpful discussions in developing this chapter, as well as the participants in the AFD Workshop Randomized Control Trials in the Field of Development: The Gold Standard Revisited (Paris, March 2019). Also thanks to Cynthia Kinnan, Jessica Goldberg, Johannes Haushaufer, Cyrus Samii, and Bruce Wydick for helpful pointers, and the three editors of the book, Florent Bédécarrats, Isabelle Guérin, and François Roubaud, for valuable comments and discussions.

5

Reducing the Knowledge Gap in Global Health Delivery

Contributions and Limitations of Randomized Controlled Trials

Andres Garchitorena, Megan B. Murray, Bethany Hedt-Gauthier, Paul E. Farmer, and Matthew H. Bonds

5.1 Background: RCTs in Medicine and Global Health

The quest for empirical, systematic, and rigorous evaluation of intervention efficacy in human populations has been a fixture of medicine long before any other discipline. One approach to rigorous assessment of treatment is the randomized controlled trial (RCT), in which investigators assign treatment status to randomly selected individuals and compare outcomes. Championed by a flourishing pharmaceutical industry, controlled trials were progressively adopted during the eighteenth and nineteenth centuries (Bothwell and Podolsky 2016). The goal was to set apart effective medical products (e.g. vaccines, antibiotics) from numerous remedies, therapies or replicas with doubtful effectiveness (Bothwell and Podolsky 2016). In the first half of the twentieth century, researchers often used alternate allocation designs (treating every other patient), but this led to important selection biases, as doctors selected patients on the basis of perceived need. Epidemiologist Austin Bradford Hill addressed this issue in 1948 when he pioneered a series of tuberculosis treatment trials that used strict concealed randomization of patients (i.e. "randomized controlled trials"). Supported by the British Medical Research Council and rapidly accepted by the research community, RCTs soon became the leading experimental design in clinical research. By 1970, the US Food and Drug Administration required the pharmaceutical industry to provide RCT results before authorizing any new drugs, giving rise to the central role of RCTs in international regulations and guidelines (Bothwell and Podolsky 2016).

The idea of RCTs as the gold standard in clinical research has been promulgated by a movement for evidence-based medicine (EBM), which aims to improve clinical practice by critically appraising the scientific literature so that clinicians

Andres Garchitorena, Megan B. Murray, Bethany Hedt-Gauthier, Paul E. Farmer, and Matthew H. Bonds, *Reducing the Knowledge Gap in Global Health Delivery: Contributions and Limitations of Randomized Controlled Trials* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/:oso/9780198865360.003.0007

can adopt best practices. Largely shaped by Archie Cochrane's 1972 book, *Effectiveness and Efficiency: Random Reflections on Health Services*, the EBM movement relies on the hierarchical ranking of the quality of efficacy studies based on the methodology used, with RCTs at the top and observational studies without a control group at the bottom. During the following decades, RCTs became an increasingly popular design outside the clinical research setting. In Western countries, their use was expanded to evaluate public policies in education, economics, sociology, and public health. The development of cluster randomized trials in the late 1970s, which randomized groups of people rather than individuals, allowed an even wider application to evaluations where individual randomization was impractical or undesirable. However, the use of RCTs for global health and international development lagged behind, with just a few dozen studies published before the 2000s (Cameron, Mishra, and Brown 2016).

A major shift occurred at the turn of the century, when the Millennium Development Goals (MDGs) were established. Recognizing that health is both a central goal of development and a potential driver (Sachs 2001), the United Nations ensured that health outcomes were prominent among them, with commitments to reduce child mortality (MDG4), improve maternal health (MDG5), and combat AIDS, Malaria and other diseases (MDG6). Funding for these areas soared (Institute for Health Metrics and Evaluation (IHME), 2016); private foundations started playing an ever-increasing role; and major organizations such as GAVI, the Vaccine Alliance and The Global Fund to Fight AIDS, Tuberculosis and Malaria were born to channel these international efforts. There arose a corresponding urgency to rigorously measure the impact of interventions in achieving those goals and to inform scale-up, which led to an evidence-based focus in global health similar to the revolution in clinical practice a few decades earlier. According to Cameron et al. (2016), 92.8 percent of all health impact evaluations from 2000 to 2012 indexed in the impact evaluation repository (International Initiative for Impact Evaluation) were RCTs, compared to an average of 66.4 percent for other fields in international development (Cameron, Mishra, and Brown 2016).

In light of the prominence of RCTs in global health, we review key contributions that RCTS have made to the field and highlight limitations that may more appropriately be addressed through other research methods in global health policy and practice.

5.2 Contributions to Policy and Practice

Since the year 2000, hundreds of RCTs have informed international guidelines for priority areas in global health. To achieve MDGs 4 and 5, evaluation efforts in maternal and child health have focused on reducing foetal growth restriction,

stunting, wasting, and micronutrient deficiencies, which are some of the largest drivers of child mortality (Bhutta et al. 2013). For example, data from such trials have informed the WHO/UNICEF recommended package of interventions for supplementation of iron, folic acid, and calcium to mothers during pregnancy, breastfeeding promotion, and the supplementation of vitamin A and zinc to children (Bhutta et al. 2013). RCTs have also provided evidence for the growth of community delivery platforms to deal with common childhood illnesses such as diarrhoea, malaria, and respiratory infections, showing that such approaches can reduce child mortality rates (Hatt et al. 2015; Whidden et al. 2018). Though MDGs 4 and 5 were not met, the scale up of RCT-tested interventions in primary care and community delivery platforms has contributed to a decrease in the under-five mortality rate by 53 percent (from 90.6 to 42.5 per 1000 live births) (You et al. 2015) and of maternal mortality rate by 43.9 percent (from 385 to 216 per 100,000 live births) (Alkema et al. 2016) between 1990 and 2015.

The majority of health-related RCTs carried out between 2000 and 2013 in low-and middle-income countries have focused on evaluating discrete biomedical interventions for either the prevention, diagnosis or treatment of a particular disease (Cameron, Mishra, and Brown 2016; Kelaher et al. 2016); i.e., more than 1,300 RCTs were conducted on HIV/AIDS (763), malaria (665), and tuberculosis (165) (Kelaher et al. 2016). These have provided strong evidence for direct HIV prevention interventions such as antiretroviral pre-exposure prophylaxis and voluntary medical male circumcision as well as new effective treatments (Krishnaratne et al. 2016). In addition, malaria RCTs have tested novel drug regimens for uncomplicated malaria, such as artemisinin-based combination therapies, as well as preventive approaches such as insecticide-treated bed nets, intermittent preventive treatment of malaria in pregnancy, or malaria prophylaxis in children (Bhutta et al. 2013; Martinez-Alonso and Ramos 2016). Community-based interventions for TB coupled with directly observed therapy have proved highly effective in improving adherence and treatment success rates (Arshad et al. 2014; The South Arican Cochrane Centre 2014), while the use of new preventive therapy regimens have been shown to be effective for TB prevention in HIV- and non-HIV infected individuals (The South Arican Cochrane Centre 2014). Current programs of mass drug administration, which are the backbone of the control and elimination strategies for many neglected tropical diseases (e.g. soil-transmitted helminthiases, lymphatic filariasis, or schistosomiasis), have been scaled internationally following RCTs (Kappagoda and Ioannidis 2014). A notable example is an RCT conducted by Kremer and Miguel in 2004, which showed significant effects of deworming drugs on school absenteeism and performance in Kenya (Miguel and Kremer 2004) and led to country-scale deworming programs around the world (Hatt et al. 2014). These represent just a few examples where RCTs have provided the global health community with a toolkit of interventions to reduce disease burdens. The scale-up of such interventions (among many other factors) contributed to a decrease of

malaria mortality by 58 percent between 1990 and 2015, of new HIV infections by 40 percent, and an estimated 37 million tuberculosis deaths were averted in the same period (United Nations 2015).

A number of influential RCTs have also provided insights into broader public health policies and reforms before national or international scale-up (Gertler 2004). A classic example is the evaluation of Mexico's PROGRESA program, which provided cash transfers to enable poor households engaging in a set of health-related activities such as prenatal care, child care, immunizations, nutrition monitoring, and educational health promotion programs. Trials that randomized these benefits to specific groups demonstrated consistent reductions in illness rates in the intervention groups compared to control populations (Gertler 2004). Since then, cash transfer programs have been tested and implemented in countries across the world (Hatt et al. 2014). RCTs have also helped inform health care reforms such as performance-based financing (PBF) schemes. An assessment of Rwanda's PBF strategy, which was initially randomized, demonstrated its impact and informed the subsequent national roll out as the country rebuilt its health system (Kruk et al. 2016). More than 20 African countries have since initiated or begun scaling PBF schemes in health care (Meessen et al. 2011). A series of RCTs have also informed the debate around the introduction of small co-payments for preventive and curative services. Conventional wisdom holds that co-payments are instrumental to promote sustainability, reduce waste, and ensure products and services are used prudently (Bates et al. 2012). However, RCTs consistently show that charging small fees for preventive products such as soap, bed nets, deworming, or water disinfectant dramatically reduces access for those who need them the most, while raising little revenue (Bates et al. 2012). Although widespread adoption of these insights has lagged, governments now provide many of these products free of charge as part of national health policies.

Despite the vast amount of evidence generated by RCTs for key areas of global health, sector-wide approaches (with cross-cutting benefits but intrinsically more complexity) are less amenable to RCT (Frieden 2017; Deaton and Cartwright 2018). The perception of the RCT as a universal gold standard design for impact evaluation and a prerequisite for the scale-up of interventions can have unintended consequences on health policy.

5.3 Unintended Consequences: Growing Gap in Evidence and Funding for Key Health Areas

Until the MDG era, there had been a relative balance between, and considerable debate over, the comparative effectiveness of vertical health interventions versus more integrated, horizontal ("system") interventions. The MDG era disrupted this balance, with most of the funding and effort channelled through an array of

vertical programs. Given the urgency and focus on a few priority areas, vertical programs were favoured as they were assumed to allow for greater service specialization, increased profile for high-priority diseases, better accountability, more rapid results, and better chance of success in weak states (Atun, Bennett, and Duran 2008). Development assistance for health grew exponentially for vertical programs targeting child health (e.g. vaccination, malnutrition), maternal health, HIV, malaria, and tuberculosis, from about 3–4 billion USD per year in 1990–2000 to more than 24 billion in 2016 (Institute for Health Metrics and Evaluation (IHME) 2016). This was backed by parallel increases in the evidence available through RCTs for the scale-up of effective interventions (previous section).

Despite their benefits, vertical programs are externally driven, top-down approaches that, without parallel investments in stronger health systems, can have negative spill-over effects such as service fragmentation, increased barriers to health care access in non-targeted populations and reduced health system effectiveness and sustainability (Atun, Bennett, and Duran 2008). Horizontal interventions such as health system strengthening and sector-wide approaches (HSS/SWAps) are complex in nature, require context-specific adaption, and act at multiple levels of a health system (Plsek and Greenhalgh 2001; Campbell et al. 2007). Evaluating them through RCTs poses considerable challenges, requires substantial investments, and in many cases is infeasible (Plsek and Greenhalgh 2001; Campbell et al. 2007). Since HSS/SWAps lacked both political commitment and RCT-based evidence on effectiveness, the percentage of development assistance for health allocated to these approaches decreased from about 15 percent in 1990 to less than 10 percent in 2016 (Figure 5.1) (Institute for Health Metrics and Evaluation (IHME) 2016). Compared with most health focus areas, gains for HSS/SWAps have not followed the predominant funding pattern. Average annual funding gains for this area dropped from 11.4 percent during the 1990–1999 period to 7.1 percent during 2000–2009 (while all other areas saw an increase in funding gains). They suffered an absolute decrease by 2.3 percent from 2010 to 2016, one of the only health focus areas to do so for this period (Institute for Health Metrics and Evaluation (IHME) 2016).

The UN post-MDGs agenda, articulated through the Sustainable Development Goals (SDGs), reflects an attempt to reduce this gap by explicitly focusing on sector-wide approaches such as universal health coverage (UHC) and HSS. The WHO has estimated that in order to achieve the health-related SDGs, nearly three quarters of all additional required investments for low- and middle-income countries in the 2015–2030 period should be allocated to HSS/SWAps, amounting to about 300 billion per year by 2030 (Stenberg et al. 2017) (Figure 5.1). Such a radical shift requires a profound rethinking of the evidence necessary and appropriate evaluation methodologies to inform funding allocation and implementation. There is growing recognition by international agencies such as the World Bank,

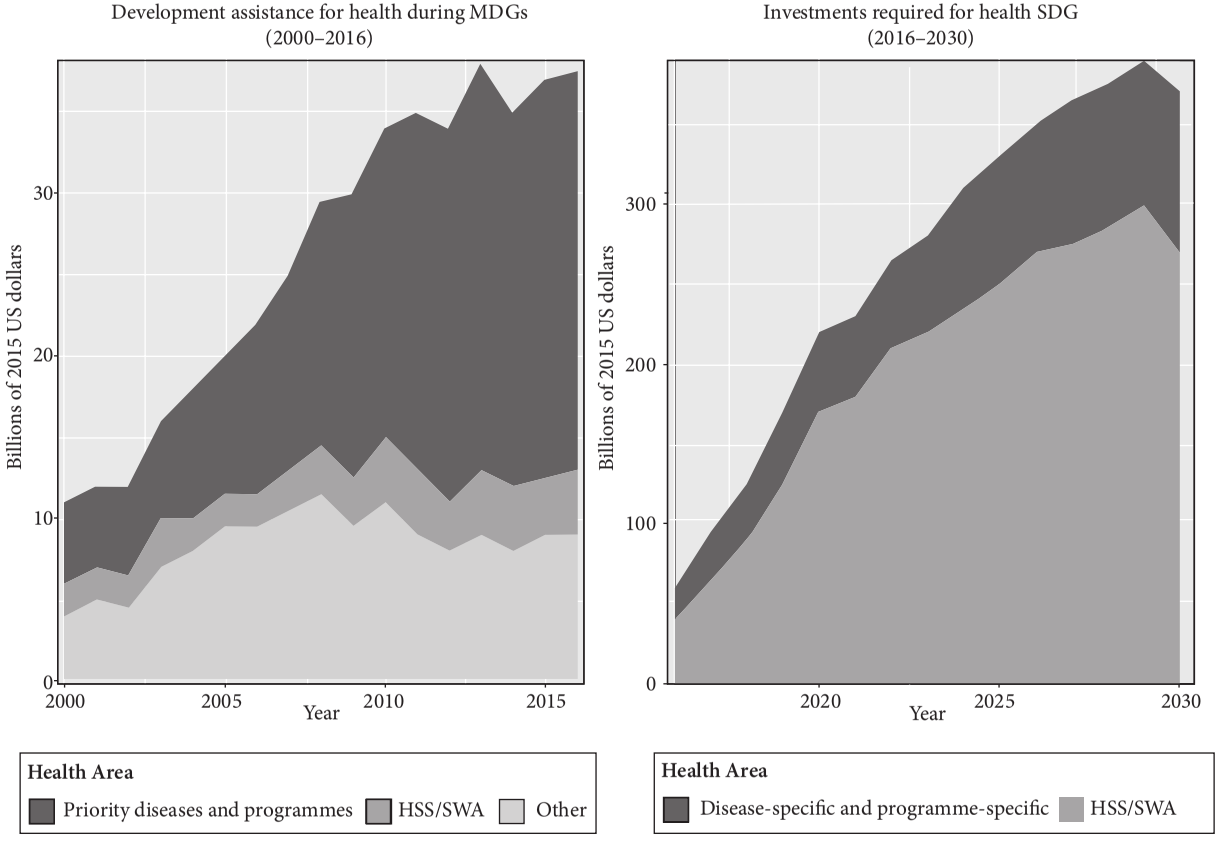

Figure 5.1 Shift in investments for health, from the Millennium Development Goals (2000–2015) to the Sustainable Development Goals period (2016–2030)

Note: Left panel shows total development assistance for health per year, adapted from the Institute for Health Metrics and Evaluation (2016). Right panel shows projected additional investments required in 67 low-income and middle-income countries to meet the health-related SDG3, adapted from Stemberg et al. (2017). It shows that a dramatic increase in HSS/SWA funding is necessary over the next fifteen years.

Source: Authors.

WHO, and USAID, among others, that the current evidence base to inform such horizontal integration is woefully inadequate despite its apparent advantages (Atun, Bennett, and Duran 2008; Giedion, Alfonso, and Díaz 2013; Hatt et al. 2015). For instance, a 2015 USAID overview of systematic reviews on HSS concluded that "additional methods for estimating the effects of HSS interventions in complex, adaptive systems are needed" (Hatt et al. 2015).

The lack of evidence for key health areas such as HSS and UHC is a sign of a broader problem in health research: the disconnect between the scope of questions that RCTs address and the type of evidence needed to improve health outcomes. An estimated 97 percent of research funding is directed at developing new health technologies (mostly pharmaceuticals) while only 3 percent is allocated to implementation research (Kruk et al. 2016). The consequence is that evaluations of effectiveness in programmatic settings and optimization of delivery are scarce, leading to substantial gaps for the scale-up of interventions under real world conditions (Kruk et al. 2016).

5.4 Challenges and Limitations

One of the most important questions in global health is why known technologies—those that are proven to work in certain settings—systematically fail to reach the people for whom they are intended. Half of the world's population lacks access to essential health services (World Health Organization and the International Bank for Reconstruction and Development/The World Bank 2017). The majority of child deaths in sub-Saharan Africa are due to illnesses—diarrhoea, malaria, pneumonia—for which solutions are known, cheap, and effective. For example, oral rehydration therapy can reduce 90 percent of diarrhoea-related child deaths globally, but only 4 in 10 children who need it receive this treatment (Kruk et al. 2016). In the majority of developing countries, Ministries of Health have set national policies based on international standards, but how best to implement those policies even at small scales remains largely unknown. The challenge is that even simple technologies require complex delivery systems—trained health workers, infrastructure, supplies, and medicines—to align at the point of care. Breakdowns occur at different scales—from individual community health workers to health care facilities or national supply chains—and are self-reinforcing (Brummitt et al. 2017), which is the basis of the movement for "health system strengthening" (HSS). There are inherent challenges in the use of RCTs to answer these fundamental questions. For an extensive review of RCTs methodological limitations, see Chapters 1 (Ravallion) and 2 (Pritchett), this volume, or the debate prompted recently by Deaton and Cartwright (Deaton and Cartwright 2018). Below, we present select topics relevant to global health practice.

The main strength of well-conducted RCTs is their strong internal validity. Under controlled conditions, they can provide an unbiased estimate of the average treatment effect for particular interventions, although critics warn that this is only true when many trial design assumptions are met (Cook 2018). Even when the results of RCTs provide an unbiased estimate of the effect for the specific population under study, this may not predict the effect of the intervention on other populations under real life conditions; i.e., there are frequently observed differences between efficacy (the impact of an intervention delivered in controlled settings) and effectiveness (its actual effect in the real world) (Ahmed, Mitchell and Hedt 2010; Shelton 2014). In contrast to observational studies, the process of enrolling participants in an RCT can artificially replace real world delivery systems to create the optimal conditions for the study, which is especially unhelpful because enrolment into a well-functioning routine delivery system is itself a central problem to solve. Furthermore, the impact of interventions as measured by RCTs conducted in different populations or contexts can vary widely. For instance, in the deworming example referenced above, meta-analyses of multiple RCTs are inconclusive about whether mass campaigns have an effect on children nutritional status, school performance or survival because of the heterogeneity of study results and differences in study inclusion criteria (Taylor-Robinson et al. 2015; Croke et al. 2016; Vrieze, 2018).

The high costs of RCTs can lead investigators to include insufficient study periods or sample sizes to properly assess the treatment effect, or to use proxy indicators that do not correlate well with or drive the outcome of interest (e.g. process indicators, narrow or early signs of illness) (Frieden 2017). For instance, removal of direct payments at the point of care through insurance schemes or user fee exemptions is a key strategy to increase health care access, provide financial protection against catastrophic spending, and ultimately improve health outcomes. While removal of user fees can affect a number of health and economic outcomes in people of all ages, an RCT in Ghana only measured malaria-related anaemia in children under five. The trial concluded that this intervention did not have measurable effects on health outcomes (Ansah et al. 2009), but the assessment was done only 6 months after the intervention started, and the study was underpowered given the low prevalence of anaemia observed (Ridde and Haddad, 2009). Design issues such as those illustrated in this example, although not exclusive of RCTs, are relatively common. A review of PubMed-indexed RCTs published in 2001 and 2006 revealed that many studies evaluated had insufficient sample sizes to detect even important treatment effects (Hopewell et al. 2010). Inadequate research designs among other issues contribute to a significant waste in biomedical research investment, which has been estimated to be about 85 percent of resources invested (200 billion in 2010) (Macleod et al. 2014).

Researchers favour RCTs because randomization can balance known and unknown factors that influence the outcome of interest, which simplifies statistical

inference in evaluation with minimal knowledge about the mechanisms behind the observed effects. However, this design can have a perverse effect of reducing information flow between researchers and the context in which they carry out the trials, which is especially important when that context involves vulnerable people operating under complex conditions. It can also result in unethical study designs (Deaton and Cartwright 2018). For example, ethical experimental trials require that the intervention is in equipoise—i.e., that there exists doubt about its benefits—but many RCTs are carried out to confirm results of observational studies while refraining from providing the benefits to individuals in the control arm (Farmer, Murray, and Hedt-Gauthier, 2013) (for details on equipoise and its implications for RCTs see Abramowicz and Szafarz, Chapter 10, this volume). Regardless of the study design, creating a body of meaningful evidence in global health delivery requires deep and long-term information loops where local actors, practitioners and implementers are actively involved in the prioritization of research questions as well as in the interpretation and dissemination of results, creating opportunities for mentored training and research that informs the services provided (Farmer, Murray, and Hedt-Gauthier 2013). In Section 5.5 we illustrate how complementary evaluation frameworks can contribute to this body of evidence.

5.5 Beyond RCTs for the Sustainable Development Goals Era: Observational Evaluation Frameworks for Health System Strengthening and Universal Health Coverage

The goal of implementation research is to know whether, how, when, and why an intervention works, and to propose further subsequent hypotheses (Bhattacharyya, Reeves, and Zwarenstein 2009; Kruk et al. 2016). It makes use of a variety of study designs, from quantitative observational and experimental methods to qualitative research (Kruk et al. 2016). There is considerable debate over the use of results from observational studies, which sometimes belie their experimental counterparts (Ioannidis et al. 2001; Prasad et al. 2013; Hemkens, Contopoulos-Ioannidis, and Ioannidis 2016; Jones and Steel 2018). However, multiple studies comparing randomized and non-randomized trials show that high-quality observational studies (e.g. prospective studies with controls) can yield comparable results to those from RCTs (Ioannidis et al. 2001; Jones and Steel 2018). The main challenge with observational studies is the greater risk of choosing inappropriate comparison groups with unmeasured factors that can bias the results. For example, an initial evaluation of the Millennium Villages Project was criticized for retroactively choosing a biased control that favoured the study results (Mitchell et al. 2018). Importantly, data and evaluation systems were initially relatively weak, without a priori control groups, which undermined the ability to draw definitive conclusions about the effectiveness of their intervention. Follow-up retrospective analyses,

however, have been more robust (Mitchell et al. 2018). In addition, there are cases where observational methods have been preferred over RCT results for decision-making despite their disagreement because they allow for longer follow-up periods, larger sample sizes and greater probability of detecting adverse effects (Frieden 2017). For instance, recommendations about influenza vaccination through nasal spray with live attenuated vaccines, which initially showed good protection in RCTs, changed over time after subsequent observational studies suggested that the external validity of the RCT findings was limited (Frieden 2017).

For sector-wide approaches, the UK Medical Research Council recognizes that designing, describing, and implementing a complex intervention is the most frequent weakness in RCTs (Campbell et al. 2000). It does not suggest alternative methodologies, but provides explicit guidelines for carrying out well-designed RCTs for complex interventions (Campbell et al. 2000). There are compelling examples where complex interventions have been evaluated through RCTs (Banerjee, Duflo, Goldberg, et al. 2015a), such as those described above for Mexico and Rwanda. However, the resources necessary for such trials limit their ability to be used widely in LMICs. Alternatively, researchers are attempting to draw lessons from the field of "complexity science" to better understand health care systems, which meet the criteria of being complex and adaptive systems (Plsek and Greenhalgh 2001). Complexity theory suggests that instead of breaking the system down to simple pieces (such as through RCTs of multiple vertical interventions), it can be better to simultaneously implement multiple approaches and gradually shift towards what works (e.g. adaptive implementation with quasi-experimental observational studies) (Plsek and Greenhalgh 2001). To achieve health-related development goals, such as for maternal and child health, what may be most important is the collective effect of an optimal suite of interventions for particular settings and populations (Shelton 2014). In this sense, one of the greatest opportunities in global health lies in adding robust data collection and evaluation methods (e.g. observational, quasi-experimental) in parallel to the myriad of health care delivery interventions taking place around the developing world. This can allow for rigorous research to be done at lower cost and without controlling the implementation process or the beneficiary population.

The limitations of RCTs in evaluating large-scale complex global health programs have led to the development of frameworks that consider an array of observational methods in parallel to program implementation, such as those proposed by the International Health Partnership (IHP+), the African Health Initiative and the Catalytic Initiative to Save a Million Lives (World Health Organization 2010; Victora et al. 2011; Bryce et al. 2013), which include major stakeholders such as the WHO, the World Bank, and the Gates Foundation. These frameworks define program success in terms of gains in intervention coverage and health effects. Studies are conducted under real-world conditions, where implementation is more variable than in controlled trials. They recognize that interventions rarely happen

in isolation, as programs from multiple agencies are implemented virtually everywhere in the developing world and changes in health and socio-economic outcomes occur regardless of existing programs (World Health Organization 2010; Victora et al. 2011; Bryce et al. 2013; El-Sadr, Philip, and Justman 2014; Reidy et al. 2018).

Using the health district as the unit of study, researchers evaluate key indicators of health system inputs, processes, and outputs (e.g. health workforce, services available) concurrently with outcome and impact indicators (e.g. coverage of services and mortality rates; Table 5.1). In addition to the continuous monitoring

Table 5.1 Indicators of coverage and mortality across the continuum of care for maternal and child health

Coverage indicators (%)	Mortality indicators
Pre-pregnancy	
Demand for family planning satisfied	
Pregnancy	
Antenatal care (at least 1 visit)	
Antenatal care (at least 4 visits)	
IPTp for malaria during pregnancy	
Neonatal tetanus protection	
Birth	
Skilled attendant at delivery	Maternal mortality (deaths per 100,000)
Postnatal	
Postnatal visit for mothers	Neonatal mortality (deaths per 1000 live births)
Postnatal visit for babies	
Early initiation of breastfeeding	
Infancy	
Exclusive breastfeeding (<6 months)	Infant mortality (deaths per 1000 live births)
Introduction of foods (6–8 months)	
DTP3 immunization	
First dose measles immunization	
Hib3 immunization	
Vitamin A supplementation (2 doses)	
Childhood	
Children sleeping under ITNs	Under five mortality (deaths per 1000 live births)
Care-seeking for symptoms of pneumonia	
First-line antimalarial treatment	
Oral rehydration salts treatment	
Improved drinking water sources	
Improved sanitation facilities	
Composite MNCH Indicators	
Composite Coverage Index (CCI)	

Source: Authors, adapted from Requejo et al. (2015).

of program implementation, additional data collection allows researchers to fill data gaps before, during, and after the evaluation period, using health-facility assessments, household surveys, longitudinal designs, and qualitative research. Quantitative analyses are complemented with qualitative descriptions of program implementation (i.e. what and how programs are implemented) and contextual factors that may have affected implementation and impact, so that results can be appropriately interpreted and lessons can be generated (Victora et al. 2011; Requejo et al. 2015; Reidy et al. 2018).

At the national level, such large-scale evaluations of effectiveness have led to the Countdown Initiative, which tracks a comprehensive list of the above-mentioned indicators for every LMIC, providing objective and robust comparisons of each country's progress (Requejo et al. 2015). At a subnational level, this framework is being used to assess the impact of complex HSS interventions, helping fill this substantial evidence gap. For illustration, consider two recent experiences in Rwanda (Thomson et al. 2018) and Madagascar (Garchitorena et al. 2018), which implement a similar set of HSS interventions, integrated across multiple levels of care (community health, primary health care centers, district hospital). Both interventions focus on improving health system readiness through horizontal programs while integrating clinical priority programs vertically. Evaluations are carried out using cross-sectional household-level Demographic and Health Survey (DHS) data obtained from representative samples of the respective populations (for Madagascar, this included a baseline in both intervention catchment and comparison area), and in repeated cross-sections at frequent intervals throughout the duration of the interventions (every five years for Rwanda and every two years for Madagascar). Intervention impact is evaluated through statistical analyses similar to difference-in-differences for a wide range of outcome indicators similar to those presented in Table 5.1, and controlling for relevant confounders (e.g. household wealth).

This quasi-experimental study design allows program managers to have the necessary authority over program implementation (when, where, and how activities are implemented), not prescribed by a research protocol. Data systems are built around the program-driven intervention so that researchers can evaluate ongoing activities and provide insights that can help managers adapt programs without interfering with implementation. For instance, a 2014–2016 analysis in Madagascar showed that despite overall improvements in most coverage indicators, access to healthcare remained very low for populations distant from health facilities (Garchitorena et al. 2018). This finding prompted an expansion of community health support, both geographically and in the scope of services provided, unfettered by research design. Moreover, additional implementation research studies allow to evaluate specific components within the overall intervention, such as a mentorship and enhanced supervision program conducted as part of the HSS intervention in Rwanda (Manzi, Mugunga, et al. 2018; Manzi, Nyirazinyoye, et

al. 2018), while helping to build research capacity among local practitioners (Hedt-Gauthier et al. 2017; Odhiambo et al. 2017).

Following the HSS intervention in one and a half districts of rural Rwanda, under-five mortality dropped by more than 60 percent between 2005 and 2010 (Thomson et al. 2018). This reduction was much higher than the rate for the rest of the country, and triple the rate needed to meet the MDGs. Similarly, under-five and neonatal mortality dropped by nearly 20 percent and 35 percent respectively in the first two years of the HSS intervention in one district of Madagascar (2014–2016), significantly faster than the average national rates observed for any country during the MDGs. Although baseline characteristics were similar in terms of per capita income and under-five mortality rates, each intervention happened in very different political and economic contexts (Bonds and Rich 2018). During 2005–2010, Rwanda experienced a virtuous cycle of political stability, international investment, and foreign aid. Madagascar, however, has been politically unstable for most of the past 50 years, with a steadily declining economy and health system investments that were the lowest in the world in 2014. Together, the two experiences provide a natural test of the extent to which integrated HSS interventions can have population-level impacts that can be replicated in different contexts (Bonds and Rich 2018).

5.6 Conclusion

The past two decades have witnessed an unprecedented improvement in health indicators in LMICs broadly connected to the Millennium Development Goals. RCTs have been instrumental in this period, facilitating the adoption of effective medical technologies and services to reduce disease burdens. Most of these services have been attempted to be scaled up vertically, which is itself amenable to RCT. Yet, large portions of the world continue to suffer from a lack of access to basic primary care. Failures in the process of scale-up are often due to breakdowns in basic implementation and weak health systems. As a result, there is a growing consensus around the central importance of sector-wide approaches such as integrated health system strengthening, integrated primary care, and universal health coverage. To achieve health-related Sustainable Development targets (SDG3), investments in these areas will need to increase more than five-fold in the next 15 years. What type of evidence should guide these investments and inform implementation?

Sector-wide approaches are less amenable to RCT, and study designs should be driven by implementation priorities. While RCTs will always be fundamental to determining solutions that are discreet and relevant in a broad range of contexts, implementation research, which typically makes use of both qualitative and quantitative methods but does not necessarily randomize implementation, can

help program managers understand how interventions with demonstrated effects can be effectively integrated into health care delivery systems. In particular, observational and quasi-experimental methods are most appropriate when the scale of the intervention makes randomization unfeasible or impractical. Adding robust data collection to the myriad of health care delivery interventions taking place around the developing world could help develop rigorous research at low costs without controlling the implementation process or the beneficiary population. Together, a comprehensive consideration of the different types of evidence available could help guide global health efforts over the next decades.

6
Trials and Tribulations

The Rise and Fall of the RCT in the WASH Sector

Dean Spears, Radu Ban, and Oliver Cumming

6.1 Introduction: The Need to Think

What is and is not gained by a focus on randomized evidence? Is anything lost? The Water, Sanitation, and Hygiene (WASH) field in international development has recently been as captivated by these questions as has been the field of development effectiveness as a whole. In this chapter, we present our participant observations from this ongoing debate. Randomized, controlled trials (RCTs) can be an important form of evidence, we conclude—just as observational studies can. But although RCTs indeed offer clear and simple computations, WASH trials have not provided clarity about conclusions. Our experience coheres with the conclusions that Deaton and Cartwright (2018) make about RCTs in development: randomization "does not relieve us of the need to think." Neither, of course, does any other type of evidence.

A core reason that thoughtfulness is required is that nobody would reasonably expect different WASH RCTs to yield the same results across different settings with different patterns of disease, demography, culture, and environment—or even testing different types of intervention. Heterogeneity of outcomes is unsurprising, given heterogeneity in inputs. The consequence is that interpreting even high-quality RCTs will always require equally high-quality reflection on non-randomized observational evidence and theoretical knowledge. Moreover, some undeniably policy-relevant questions will never be suitable for randomized experimentation: sewer lines, religion, and population density will not be randomized (although treatments that interact with these might). Even questions that could, in principle, receive randomized evidence, in practice are unlikely to within the decade that remains before the sanitation Sustainable Development Goal is intended to be accomplished.

Throughout, we urge the reader to compare and contrast an RCT of a WASH intervention in a rural part of a developing country with the paradigm case for an RCT: a clinical drug trial in a research hospital. In repeated trials of the same drug, researchers indeed test chemically *the same drug*. Ultimately, if a set of drug

Dean Spears, Radu Ban, and Oliver Cumming, *Trials and Tribulations: The Rise and Fall of the RCT in the WASH Sector* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0008

trials are implemented comparably, they provide multiple estimates of the same causal parameter, suitable for a meta-analysis. The same chemical does not work the exact same way in each body, but the outcomes across studies are likely to be drawn from the same distribution. Even if the clinical trial populations differ, theories and evidence from the medical literature are likely to be rich enough to point toward explanations. WASH trials, in contrast, are bound to be few and unlike.

In this chapter, we do not have the space to cover every aspect of WASH that is important in developing countries. Urban infrastructure, such as sewerage or fecal sludge treatment, would be essentially impossible to evaluate in a randomized study, but is still believed to be an important factor in public health, based in part on demographic literature (Cutler and Miller 2005[1]). An important literature in WASH, that we do not address, considers the role of blinding in household-level RCTs. In hand hygiene, for example, whether soap is antibacterial or not can be blinded, but whether household members are encouraged to wash their hands cannot. An influential review in household water treatment investigated the role of incorporating blinding into the physical design of water filters. For the rest of this chapter, we set these issues aside to focus on *rural sanitation* in developing countries—not because this is the only important dimension of WASH, but because it is a useful case to understand what RCTs can offer.

6.1.1 Background: Recent Evidence in Sanitation and Child Health

The rural sanitation sector is facing confusion and debate. Ambiguity and contestation themselves are not surprising. What is surprising in this case is that disagreement and perplexity have resulted substantially *as a consequence of recent RCTs*—even though what RCTs are supposed to bring is simple clarity. Three recent sanitation RCTs (Null et al. 2018, Luby et al. 2018, and Humphrey et al. 2019) have particularly led to heated debate both in the scientific circles and in the implementing sanitation sector organizations. A consensus paper (Cumming et al. 2019) lays out a unified interpretation of the results by a multidisciplinary group of researchers. As we discuss in more detail in Section 6.2.2, what these three trials have in common is that they all find null results regarding the impact of improved sanitation on child height-for-age. What makes this outcome even more apparently paradoxical is that many observers interpret the three recent RCTs in question to have yielded essentially the same answer (whether or not that is the right conclusion to draw is part of the contestation).

[1] While this does not materially change our argument, it should be noted that Anderson, Charles,, and Rees (2018) re-evaluate the findings from Cutler and Miller and find the effects of water filtration to be significantly smaller than in the original paper.

Moreover, nobody argues that the trials were not carefully conducted, or that they did not successfully implement what they had intended to do.

What went wrong? Although the WASH sector's current predicament is not what is imagined by proponents of RCTs, it does match some of the warnings by Cartwright and Deaton. In particular, although there is widespread agreement on *what* treatment effect the RCTs estimated, there is widespread disagreement about *why*. There are more candidate explanations than there are RCT data points. This question—why—is exactly what is needed to know for policy-making: the ability to understand what factors are important for next time. But, as Cartwright and Deaton caution, "why?" is exactly the question that RCTs can be challenged to answer.

In this chapter, we discuss reasons not to be surprised by this seemingly paradoxical outcome: WASH in development in fact has many properties that make effects heterogenous and context-dependent. So, no small set of studies can resolve its challenges—a key difficulty if studies cost tens of millions of dollars and take years to complete. But before we can draw such general lessons, in this section we introduce the evidence from these recent studies and others.

6.1.2 Lessons from a Range of Studies

Exposure to poor sanitation is widely suspected to be bad for children's health. It strikes most people as intuitively plausible that a child who is routinely exposed to feces can be expected to do worse, on average, than one who is not. But how much worse, and which remedies work well enough? Part of measuring how much worse is deciding on an outcome indicator. Arguably the most important indicator is early-life mortality, but binary outcomes have low statistical power, and fortunately, even in the most deprived populations, most children do not die. The result is that studying mortality requires very large samples—too large for a plausible RCT. Another option is diarrhoea: researchers have interviewed mothers about the looseness or firmness of children's stools in hundreds of surveys. The challenge here is that a loose stool is a subjective judgment: in India, better-educated mothers report more diarrhoea, presumably because they are more likely to see a given situation as a problem.

Over the course of the last decade, a statistically ideal outcome-variable has emerged. It is a continuous variable, objectively measured, and a property that everyone possesses: anthropometry, and especially child height. At about the same time that WASH scholars were prominently hypothesizaing height as an important variable influenced by exposure to poor sanitation (Humphrey 2009), economists were discovering the importance of height as a component of economic human capital—because of what height reveals, at the population level, about early life health, disease, and net nutrition (Case and Paxson 2008). In

some countries—such as during India's political movement for a "right to food"—child height statistics even rose to prominent attention in public policy debates as a measure of nutritional outcomes and deprivation. Policy actors associated with left and right parties in India debated whether the short average height of Indian children reflected unproblematic genetics, urgent food shortfalls, women's status—or widespread exposure to open defecation (Coffey et al. 2013).

So, a set of papers have emerged studying the effect of various dimensions of sanitation, especially on height. The exact notions of sanitation vary considerably: some look at open defection without using a toilet or latrine; some investigate the use of improved latrines, rather than simpler ones; and some consider community-level averages of these to be important, rather than only one household's behavior. These roughly sort into four groups by methodology (RCT or observational) and by quality (higher and lower impact), resulting in high- and low-quality RCT and observational studies.

Low-impact RCTs. An initial group of RCTs did not have as much of an impact on sector thinking as the more recent ones did. In some cases, the implementation of the RCT suffered from irregularities or incompleteness that limited what could be learnt. Several of these were organized by the World Bank Water and Sanitation Programme, in several countries. One, in Maharashtra, was not ultimately conducted by the state government in the full set of districts on which the World Bank had planned (Hammer and Spears 2016). In another, in Madhya Pradesh, the sanitation treatment was implemented so much later than intended that only a few weeks remained before the endline survey (Patil et al. 2014). Another RCT, in Orissa, was conducted with the highest standards of care and rigor, but did not have the opportunity to learn much about the effect of open defecation on child outcomes, because the intervention did not, in the end, generate a large difference in open defecation between the treatment group and the control group (Clasen et al. 2014). As we will discuss, social forces in rural India make the promotion of latrine use, rather than open defecation, possibly more different than in other developing country contexts—but in other contexts open defecation is becoming ever more uncommon.

Low-impact observational studies. Perhaps the largest set of papers are the many observational papers which shed little credible light, especially about what the causal effect of sanitation may be. In many cases, of course, the goal of a study was simply to describe an important situation or pattern in the data; it is no criticism of these that they do not achieve what an RCT is intended to achieve. However, many of these are too quick to draw causal conclusions from comparisons—perhaps with regression controls—without any special reason to conclude that the situation is one in which correlations are likely to informative about causation. Some of these are as simple as comparing height among children who live in households with a toilet or latrine against children who live in a household without one. We do not know of any case where such a household-level

observational sanitation comparison offers a credible estimate of a causal effect—not least because it ignores the spillover effects of some households in a locality on their neighbors.

Many observational researchers are therefore careful enough not to draw causal inferences from such comparisons, but some are not. What matters here is what the authors claim: whether they are clear and careful about the questions they ask, or whether they overreach in describing what their data tell them (either in the study, or in broader policy conversations).

High-impact observational studies. A third group of studies is also observational, but more carefully considers patterns of evidence from which more informative conclusions can be drawn. Many of these are inspired by what some empirical researchers in econometrics have called the "credibility revolution" of causal identification studies. Yet careful methods for investigating, doubting, and double-checking evidence of impact from observational data long predate the last few decades of econometrics. Indeed some, such as social statistician David Freedman (1991), pinpoint the origin at John Snow's study of WASH and cholera in London. Demographers and epidemiologists have all been part of the best of this literature, along with econometricians.

What distinguishes some of these studies is that they look for special cases that can be learnt from. Usually, this involves rich, context-specific understanding of why factors that are likely to be confounding in other cases are unlikely to be a problem in the case under consideration. For example, in the original John Snow (1855) study, he argues that cholera is transmitted through the fecal contamination of water supply. His argument is based on the much higher cholera mortality rates among households supplied by the Southwark and Vauxhall water company, whose intake from river Thames was downstream of the sewage discharge point, relative to those supplied by the Lambeth water company, whose intake was upstream of the sewage discharge point. In making this argument, John Snow pays such close attention to confounding factors that his words bear repeating.

> The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other [. . .] As there is no difference whatever, either in the houses or people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.

In other words, credible conclusions are not the result of a math trick or technical procedure; econometricians who present their field this way have failed to grasp that what advances their science is discernment between the cases that one can

probably learn from, and the cases where evidence suggests one cannot. Instead, credible evidence is often rooted in social scientific understanding of a situation, verified with a long series of empirical verifications. As a result, many of these studies are longer and more detailed than computationally simpler studies in the literature which make use of random assignment.

6.1.3 Recent Evidence from High-quality RCTs

A fourth category contains high-quality RCTs. Within the last year, results have emerged from three RCTs in three developing countries, each designed to measure an effect of a type of rural sanitation intervention on child height and other outcomes. Two of these RCTs are part of the multi-site WASH-Benefits (from here on referred to as WASH-B) trial: one from Kenya (Null et al. 2018) and one from Bangladesh (Luby et al. 2018). The third reported on the SHINE trial in Zimbabwe (Humphrey et al. 2019. Although there were differences among the three studies, each of them focused on early-life health and nutrition of children in the rural part of a developing country. Each of them upgraded the sanitation of a young child's household or compound[2] and promoted latrine use and hygiene behavior. Each of them compared a sanitation intervention to a nutrition (here, meaning feeding) intervention. And ultimately, none of the three showed an effect of their sanitation treatment on child height.

In the many months since the results of these three high-quality RCTs were released, debates have emerged about what conclusions can be drawn from them (Cumming and Curtis 2018). Before even considering *why* the three RCTs found the result that they did, researchers have debated whether the findings are surprising or are expected. Coffey and Spears (2018), for example, show that Demographic and Health Surveys from the rural parts of the three countries show no association between child height and community averages of the variables tested, with only the simplest of controls for socioeconomic status.

Arnold et al. (2018), in contrast, point to the large number of poorly structured, household-level observational studies as evidence for a claim that observational studies in the literature predicted that these RCTs would find statistically significant effects. In particular, Arnold et al. show that children in the baseline dataset from the same sites are taller, on average, in households that have latrines than in households that do not. Based on their comparison, they conclude that, because their RCT did not show an effect, this contrast offers a reason to doubt all observational evidence in general, even evidence using careful strategies, and even evidence asking different questions. We read this as a straw man argument: *of course* it is

[2] In WASH-B the sanitation intervention provided improved toilets to the entire compound (ranging in size, in Bangladesh, from 3 to 10 households and in Kenya from 1 to 4 households). In SHINE the sanitation intervention provided improved toilets to the child's household only.

possible to construct an observational study that is not credible, as Arnold et al. did, just as it is possible to conduct a sloppy RCT that proves uninformative. But everyone's understanding will only progress if we consider the best evidence available from each approach.

In the search for explanations, it has become clear that there are more candidate explanations than there are studies to adjudicate amongst them. RCTs alone provide little guidance on how to *use* their estimates in further science and policy-making. Deaton and Cartwright merit quotation in full:

> The results [of RCTs] cannot be used to help make predictions beyond the trial sample without more structure, without more prior information, and without having some idea of what makes treatment effects vary from place to place or time to time. There is no option but to commit to some causal structure if we are to know how to use RCT evidence out of the original context. Simple generalization and simple extrapolation do not cut the mustard. This is true of any study, experimental or observational. But observational studies are familiar with, and routinely work with, the sort of assumptions that RCTs claim to (but do not) avoid, so that if the aim is to use empirical evidence, any credibility advantage that RCTs have in estimation is no longer operative. And because RCTs tell us so little about why results happen, they have a disadvantage over studies that use a wider range of prior information and data to help nail down mechanisms.

In the case of SHINE and WASH-B, the lead authors of the studies have proposed in an integrative review paper (Pickering et al. 2019) that their particular interventions did not show an effect because households were insufficiently often treated with behavior change encouragement. This is possible. Another possibility is that sanitation effects occur from village-level externalities (Geruso and Spears 2018), rather than own-household-level changes: one child's household is a small part of the child's entire sanitary environment. Andrés et al. (2017) find that a child in rural India who moves from a household without improved sanitation in a village with low take-up to a household with an improved sanitation system in a village with high take-up can see a reduction in diarrhea prevalence of 47 percent. Of this reduction, a quarter can be attributed to the direct benefit of access to sanitation, with the remaining three quarters accrues from the indirect effect of neighbors utilizing improved sanitation. The relationship between the share of the village with improved sanitation and diarrheal prevalence appears nonlinear. There is almost no externality in villages with low sanitation takeup. The best available observational evidence finds effects of such context, community, neighborhood, or village-level changes, not of comparisons between one household with safer sanitation and neighboring household without. It is important to call out that in the above-mentioned RCTs (SHINE and WASH-B) the size of the intervention cluster was typically lower than would be expected if

externalities were fully taken into account. In most sanitation RCTs the size of the intervention cluster is a village. In the SHINE trial, only the child's own household was treated; in the WASH-B trial, only compounds were treated, generally amounting to about two households, counting the child's. These represent much smaller changes in the child's environment than have been hypothesized in some of the existing literature to be important.

Another set of possibilities is that the effects of sanitation are heterogeneous—so it is inappropriate to call them effects of simply "sanitation" at all (Cumming and Curtis 2018). For example, some have argued that it is open defecation in particular that is harmful to children's health, but in the rural Kenyan and Bangladeshi settings of the WASH-B trials, open defecation was already very low at baseline: single digits of percentage points.

Yet another heterogeneity that has been hypothesized in the literature concerns population density. As a predictor of child height and infant mortality, local open defection rates interact with population density: open defecation is more steeply associated with lower child height in places where population density is high (Hathi et al. 2017). In places where population density is low, observational data does not show an association, on average. In places where population density is as low as rural Zimbabwe and Kenya, Hathi et al. show that observational data do not predict any association between child height and local sanitation at all—just as these trials found.

Figure 6.1 illustrates that observational data does not predict an association between sanitation and child height in every context. The figure is computed with Demographic and Health Survey data[3] from rural Zimbabwe, the site of the SHINE trial. The horizontal axis is chosen to match the type of sanitation coverage promoted by the experiment: "improved" sanitation, which means neither open defecation nor the use of unimproved, simple latrines. In particular, sanitation is averaged among all households in the village, so a value of 0.4 indicates that 40 percent of randomly sampled households report using unimproved sanitation. The vertical axis is the measure of child height-for-age that is common throughout this literature. Each dot is a rural village. The perfectly flat association—matching Hathi et al.'s predication for a context of such low population density—demonstrates that, in rural Zimbabwe, the highest-quality, population-level observational data show no association between height-for-age and this measure of sanitation. In contrast, the same Demographic and Health Survey data show steep associations between local open defecation and child height in densely populated India and other contexts. Figure 6.1, of course, cannot provide

[3] Observations are all rural children under 60 months old in the birth recode of Zimbabwe's 2015 Demographic and Health Survey. Height-for-age is as computed by the DHS, according to the 2006 WHO world reference standards. Each child is matched to average household sanitation computed for all households in the household recode, in its primary sampling unit (PSU). Dots are PSU averages. The line, computed at the child level, includes a 95 percent confidence interval.

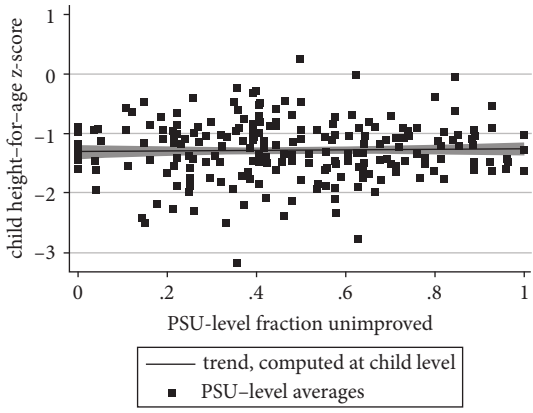

Figure 6.1 Association between improved sanitation and child height-for-age in Zimbabwe

Source: Authors. Details in the text.

a full explanation of any study's results; its contribution is to rebut the claims of some RCT proponents that the recent evidence is at odds with observational data, or that observational data predicted that the SHINE trial would be likely to show an effect. Correlation is not causation, but observational data for Zimbabwe show no correlation.

Our point is not that we know which of these possibilities explains the SHINE and WASH-B results. To the contrary: nobody can be sure. Our point is that such a predicament is unsurprising. Indeed, the challenges for sanitation more broadly or WASH as a sector are even deeper than these. There are many types of "sanitation" interventions. It is also important to note that, as argued in the consensus paper (Cumming et al. 2019), these three studies do not challenge the evidence (from high-impact observational studies, such Cutler and Miller (2005)) that large-scale improvements in water and sanitation played an important role in improving child health in the growth of today's high income countries.

6.2 The Gold Standard? Challenges for Sanitation RCTs

A paradigmatic case of a health RCT is a clinical drug trial: out of a group of similar patients, some are randomly selected to receive a pill, others are randomly selected to receive a placebo. Swallowing a pill is sufficient for it to be fully administered; the pills received by the treatment group are chemically identical to any other instance of that drug. Perhaps this is indeed the "gold standard" for medical research—although Deaton and Cartwright challenge this interpretation, too. In this section, we show that such a promising metaphor does not apply to sanitation in rural developing countries.

6.2.1 Parameter Heterogeneity: Different WASH RCTs *Should* Give Different Answers

Some of the highest-cited papers in medicine and epidemiology are meta-analyses that combine multiple studies. Recognizable for their "forest plots" of stacked confidence intervals, they combine estimates from many individual studies to compute a pooled estimate of an effect size and confidence interval. A core benefit of computing a pooled estimate with a forest plot is to reduce uncertainty from sampling error. Even if each individual study's sample is small, so its confidence interval is large, the pooled estimate can have a small confidence interval. In other words, several imprecise estimates of the same quantity can, together, provide a precise estimate of that quantity.

Again, the paradigmatic case is a set of clinical trials of a single, chemically identical drug, on comparable populations. In this case, the several RCTs indeed are multiple estimates of the same quantity. WASH studies—randomized or otherwise—are almost never so uniform in the questions that they ask. Parameter heterogeneity is not itself a criticism of RCTs.[4] However, *in practice*, RCTs will be challenged to appropriately handle heterogeneity in effects if studies are constrained to be few, expensive, and time-consuming to create. Moreover, a feasible rural sanitation RCT must be limited to one or a few settings, unlike observational studies which can be representative of the heterogeneity across entire populations, or even sets of countries. Here we review some example heterogeneities:

Type of WASH. One Sustainable Development Goal, for example, is to end open defecation; another is to provide safe and affordable access to clean water. Both of these are important targets, but information about how to achieve one of them, or what the effects of achieving it may be, is not directly relevant to the other. More broadly, WASH includes urban sewage networks and treatment systems; hand hygiene (with or without soap, with or without anti microbial agents); water treatment (at the source, at home; physically filtering, chemically, or with the sun; blinded or unblinded); toilets (unimproved latrines, improved latrines, and so on); and more.

Type of sanitation. Differentiating among open defecation, unimproved latrine use, and improved latrine use is a subset of differentiating within WASH, but we highlight it because it is so often overlooked. The WASH-B studies, for example, were conducted in contexts that were essentially already without open defecation before the experiment started. So, they cannot directly speak to the benefits of reducing open defecation in the contexts where it remains common, such as rural India.

[4] By heterogeneity we mean differences in local conditions, and not differences in individual responses to interventions.

When the WASH-B studies were published, *The Hindu* (a leading Indian newspaper) published an article with the headline "Link between sanitation, stunting questioned" (Pulla, 2018). The article was motivated in the opening paragraph with reference to open defecation, a high-profile policy issue in present-day India. But, the article never explained that the WASH-B studies—on which the *Hindu* article focused—were not about open defecation at all; that Bangladesh is essentially open defecation free; or that in rural India, in contrast, most rural households defecated in the open at the time.

To be sure, leading researchers, themselves, would be unlikely to conflate different categories of sanitation. Moreover, newspaper reporters could be just as likely to misreport randomized and non-randomized evidence. Few researchers, of any methodology, are enthusiastic to correct a newspaper reporter in the exciting moment when a study is finally generating public attention. However, because RCTs are rare and costly (in money and time), they generate an impulse publicly to proclaim an importance that might extend beyond the question to which the study actually speaks.

Community or own-household effect. It may be the case that one household's children are impacted by the germs introduced into the environment by other households' sanitation behavior. If so, switching only a child's own household away from open defecation, for example, might not produce a large difference in her sanitary environment. There may be effects of both own-household sanitation and of community-level sanitation coverage.

Population density, urban/rural context, and other environments. The same difference in exposure to sanitation might represent a larger or smaller difference in exposure to fecal pathogens, in contexts where people live nearer to or further from one another. As we have discussed, for example, Hathi et al. (2017) finds that open defecation interacts with population density to predict infant mortality and child height.

Background health and health behaviors of the population. WASH interventions may have different effects depending on the baseline health of the population. Deworming interventions, for example, will do little good in a context without worm infections. Clean water has been shown to interact with breastfeeding.

Perhaps most relevant in our context is that many studies and policy accounts conceptualize child height in terms of *stunting*, rather than height as a continuous variable. Stunting is a dichotomized measure of height that separates children into much-too-short and otherwise. Dichotomizaing height in this way reduces the statistical power of a study to find an effect, even if there is one (Spears et al. 2013). Moreover, it can be a source of parameter heterogeneity: a sanitation intervention could make average height-for-age taller by the same amount in two different populations, but have much different effects on stunting depending on whether the average child is near or far from the stunted boundary.

Cultural and social factors. The same sanitation intervention might have different consequences if implemented in different societies. This possibility has been especially well documented in South Asia, where caste (Lamba and Spears 2013) and religion (Ghosh, Gupta, and Spears 2014) have both been seen to predict and, more broadly, to shape open defecation practices. Muslims in India, for example, are poorer and more disadvantaged in public services, on average, than Hindus in India, but are less likely to defecate in the open. Because of residential segregation, Muslim children are more likely to live near other Muslim children, and vice versa. The consequence is that Muslim children are more likely than Hindu children to survive the first year of life; prior to the explanation based on community-level sanitation, demographers called this puzzle the "Muslim mortality paradox" (Geruso and Spears 2018). Other factors are likely to be important, too. Our observation is that the differences in this list have undeniable scientific and policy relevance. There are at least two challenging implications for using WASH RCTs for evidence-based policy. One concerns the cost-benefit ratio of investing in a randomized study, instead of other sources of evidence. Perhaps an RCT, if implemented successfully, will have advantages in the clarity of conclusions about cause and effect (although see Deaton and Cartwright before being sure). However, the set of cases that the RCT speaks to may be smaller than in the paradigmatic case of a clinical drug trial, because of all of these sources of heterogeneity. The other challenge is that incorporating such heterogeneity into a study is the exact sort of thing that a high-quality observational study can do, by exploiting large samples in demographic data, and variation across contexts. Indeed, cultural contexts, environmental backgrounds, population density, and the baseline distribution of height are all factors that cannot be randomized, and must be understood through observational studies or observational complements to an intervention study.

6.2.2 Type 3 Errors, Weak First Stages, and Treatments that Do Not Treat

Observational studies attempt to learn from variation that already exists in the world. The advantage is that cases are available to compare spanning the range of good and bad sanitation, often within a single country. In other words, big differences in exposure to sanitation already exist. Challenges, of course, can remain, if variation in WASH exposure is correlated with variation in other dimensions of advantage or disadvantage—which is why the best observational studies look for special cases to learn from, and examine them closely. Randomized interventions studies attempt to *generate* variation that can be learnt from. This can be a challenge if variation is difficult to generate.

Consider an extreme example. In actual development policy, almost everyone agrees that rapidly eliminating open defecation is a high policy priority. But imagine if that were not the case, and if policy-makers instead were trying to decide whether to consider the elimination of open defecation to be a priority, on the basis of its consequences for health. Further imagine that the policy-makers agreed that the only admissible evidence would be from RCTs. Finally, suppose that (because eliminating open defecation was not already a policy priority) nobody yet knew any techniques that were likely to successfully reduce open defecation. In this situation, it could be the case that open defecation has very large health consequences in some contexts, but no admissible evidence could demonstrate the health benefits convincingly enough to justify investing in learning how to generate variation in open defecation.

Fortunately, WASH is not quite in that paradoxical situation. However, the case of open defecation, in particular, has suffered from difficulty in generating variation in exposure to open defecation—especially in exactly those contexts where the effect might be largest. Because of the history of the caste system and the continuing importance of untouchability, rural India has proven particularly resistant to change in open defecation (Coffey and Spears 2017). Clasen et al. (2014) conducted a careful and high-quality study of open defecation in rural Odisha, a poor state in India. Unfortunately, as the authors conclude in the study, the intervention did not result in a large change in open defecation behavior, and therefore did not generate enough variation in exposure to open defecation to learn from. However, if it is the case that open defecation is more harmful to child health where population density is high, then the inability to generate RCT-style evidence *from rural India* will have the consequence that the literature will overlook what may be the largest effect, and in the context where open defecation remains most common. It is not logically necessary that the places where problems persist will tend to be the places where policy-makers do not have the tools to solve them—but it would not be surprising if this often turns out to be the case.

So, an intervention study cannot learn if it *attempts* to generate a difference but then does not. One type of this problem is exemplified by the Clasen, et al. study: study participants do not take up the treatment, and do not change their behavior. In this case, researchers may learn something useful about what does not change behavior, but they will not learn about health effects. Another instance of this problem can occur even when the study is implemented exactly as planned and behavior change occurs exactly as hoped, if the study is designed at the wrong level of treatment. For example, consider the SHINE and WASH-B trials which treated only one or two households in a child's village. If the relevant factor in determining exposure is community-level sanitation, then these interventions would not have generated large differences in exposure, even if implemented perfectly.

In collaborating across disciplines to write this chapter, we learnt that this phenomenon has a variety of names. The medical and epidemiological literatures sometimes use the term "Type 3 error" (an analogy to familiar inferential Type 1 and Type 2 errors) to describe an intervention that effectively did not happen. The error in this case would be to conclude anything about what the effect of the originally proposed intervention *would be*, if it happened. The economics literature describes the same phenomenon as a "weak first stage." In many studies, the randomized treatment is just a tool—or an "instrument" in this language—to learn about the effect of a first-stage variable on a second-stage outcome. For example, a randomized treatment might attempt to generate differences in exposure to open defecation (the first stage), in order to learn the effect of open defecation on child height (the second stage). If the first stage is not strong or statistically clear enough (there are formal tests for a weak first stage), then the study will not be informative about the second stage effect of interest.

Conceptually, the point is simple: *If an intervention study does not or cannot generate a large difference in exposure to a type of sanitation, then it will not be very informative (in favor of or against an effect) about the effect of such sanitation*. Statisticians complicate that simple point by writing about "power calculations": the *power* of a study is the probability of detecting an effect, if there indeed is one. A study with a small sample and with a weak first stage is likely to have low power. The consequence would be a wide confidence interval for the final effect estimate. A wide confidence interval is just a fancy way of saying that the study did not learn much: it cannot rule out large effects, or no effects at all. So, for example, when an instrumental variables method is used to produce a confidence interval for the effect of local open defecation on child height, using the Clasen et al. (2014) Orissa data, the result is a very large confidence interval. The confidence interval includes the possibility of very large effects of open defecation on child height, the possibility that there is no effect whatsoever, and even the possibility of perverse effects in which open defecation makes children better off. A wide confidence interval is simply statisticians' way of saying that not much was learned.

We can estimate the power of sanitation experiments that have not happened by making plausible assumptions and using existing demographic data. For example, Geruso and Spears (2018) compute the sample size that would be required to estimate the effect of open defecation externalities on infant mortality. Infant mortality is an outcome variable that is at least as important as child height, but it is more difficult to study in a moderately large sample because it is dichotomized and because infant deaths are too common, but statistically rare. An adequately powered study would require a very large sample. Geruso and Spears calculate that such an intervention, in their Indian setting, would cost about ninety million dollars, excluding any costs of data collection, management, and analysis, assuming optimistic first-stage effects on sanitation behavior.

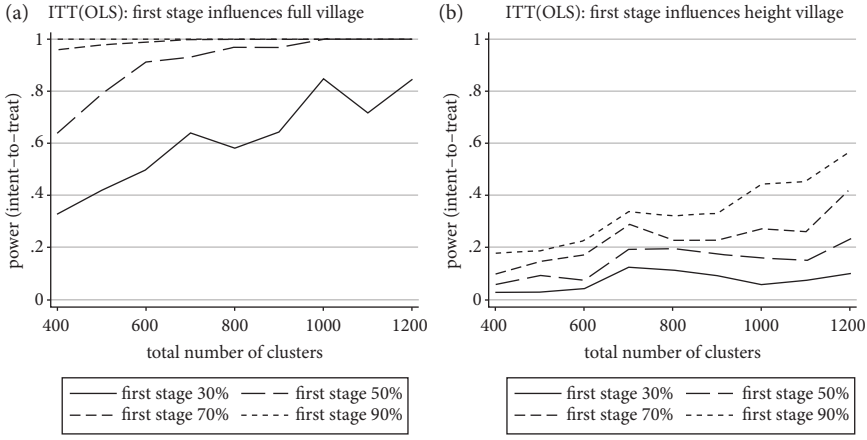

Figure 6.2 Monte Carlo simulations of power of hypothetical sanitation experiments in rural India, under various assumptions about the first stage effect on village open defecation

Source: Authors. Details in the text.

In Figure 6.2 we make similar computations for hypothetical experiments to study the effect of open defecation externalities on child height. Arnold et al. (2011) describe in detail a basic strategy for estimating the statistical power of a study using randomized simulations. The figure uses India's 2005–6 Demographic and Health Survey. The simulations behind the figure assume that there is a large, true, uniform, constant, linear effect of exposure to open defecation: 0.5 height-for-age standard deviations, linearly associated with moving from 100 percent to 0 percent of a child's village defecating in the open. The assumptions about the "first stage" of the study vary in two ways. First, the intervention is assumed to have one of four strengths of effect on open defecation behavior, reducing open defecation in treated households by 30 percent, 50 percent, 70 percent, and 90 percent, represented by the four lines within each panel. Second, in panel (a) all households in intervention villages were treated with the intervention, but in panel (b) only households with a child with measured height were treated. Thus, in panel (b) only a smaller fraction of the households are assumed to be treated (although still a larger fraction than in the SHINE and WASH-B trials).

In each panel, the horizontal axis is the assumed number of villages; the axis *starts* at 400 villages, which would be a large experiment by the standards in the literature. The vertical axis is the fraction of simulated RCTs in which a statistically significant effect size was detected. In all cases, the data are constructed assuming a large, constant effect, so the vertical axis reports the probability of seeing an effect that truly is present.

Researchers sometimes use 80 percent as a reasonable level of power for an experiment to detect an effect that is truly present. In panel (a), where all

households in a village are treated, studies can reach this threshold, if the sample size is large enough, if the first-stage effect is large enough, or both. Panel (b) shows that even very large sample sizes—much larger than could plausibly be funded—with large first-stage effects do not show adequate power, if only the households with children are treated by the sanitation intervention. Of course, these simulations, like any, reflect their assumptions: in this case, the critical assumption is the large effect size on village-level, rather than household-level, sanitation. Under these assumptions, Figure 6.2 computes that the power of even a very expensive RCT to detect an effect depends critically on the strength and nature of the first stage.

6.2.3 Important Questions that Will Not Be Randomized

In this chapter, we have emphasized that there are higher- and lower-quality studies of every methodology, and that RCTs and observational studies have their own, non-overlapping set of challenges. Here we wish to respond to an extreme argument that *all* non-RCT evidence should be mistrusted. The following are some example questions that are important for urban and rural sanitation policy in developing countries, and that will never be answered only with an RCT:

- What are the health effects of increased usage of urban "safely managed[5]" sanitation (either through upgrading a city's sewer and sewage treatment facilities, or through improving non-sewered management of fecal sludge)?
- More broadly, what are the consequences of a city-wide inclusive sanitation system, which cannot exist merely in part?
- Are Muslims in India more likely to use latrines than Hindus? Why?
- Is the effect of open defecation different in cities, or in high population-density places?
- What is the effect of open defecation on infant mortality—a rare, dichotomized dependent variable that is very important, but realistically impossible to study with an RCT, due to statistical power?
- What are the long-term effects of safe sanitation over the generations, so children are born to mothers whose own uterine environment was free from nutritional loss due to fecal pathogens?
- How much open defecation is found within households that own a latrine, even a latrine that some people use? Where? Why?
- Is lack of access to open defecation a particular challenge among the poor, women, or the elderly? Are the costs of inadequate sanitation greater for these groups?

[5] As defined under the SDG 6.2 indicator.

RCTs alone cannot answer these questions either because they are not ultimately about cause and effect, or because they concern a factor (such as culture or place) that cannot be experimentally varied. Furthermore, and in particular in urban areas (vs. rural areas), sanitation is not just a household intervention of building and using a latrine, but rather a system of managing the flow of human waste (through underground sewers or through non-networked emptying, transporting and treatment). Such a system includes household behavior but also infrastructure and regulation and the last two do not lend themselves to experimental manipulation.[6] RCTs can contribute to answering many of these questions, but not alone. In development, the "evidence-based policy movement" has proven to be almost synonymous with the push for RCTs in policy-making. But many policy-makers lack even evidence describing where and among whom challenges are found: how much open defecation, for example, is found in various Indian districts. Other questions are necessarily about interactions. If population density or cultural practices indeed interacts with a treatment to shape its effectiveness, then a policy-maker (such as a World Bank or Gates Foundation decision-maker) needs to understand that interaction, which necessarily requires understanding observational, non-randomized variables.

Even the standard set of health questions may take more time to answer than is commonly recognized. For example, one possibility is that there are intergenerational pathways by which net malnutrition is transferred: a mother who is unhealthy in childhood may grow up to provide a smaller uterine environment, which impacts child growth. Alternatively, exposure to disease as an adult—if it reduces a mother's pre-pregnancy body mass or weight gain in pregnancy—could leave a mother's body less able to nourish a child in utero and during breastfeeding; this would be true even if a hypothetical intervention study fully eliminated fecal pathogens immediately before a child was conceived. Observation studies, in contrast, can learn from long-term equilibrium differences in exposure to sanitation.

6.3.4 The Overlooked Issues that Could Be Advanced with RCTs (or Non-RCT Intervention Studies)

Finally, we note that some public health questions could be advanced with more and different RCTs: questions about how to change the behavior of individuals or other relevant economic agents. In the field of development economics Blattman (2008) refers to these types of RCTs as "Impact Evaluation 2.0." He contrasts the

[6] While this is beyond the scope of this chapter, we note that the effectiveness of infrastructure (be it sanitation, roads, public, healthcare) development is critical for poverty alleviation and yet it is inherently outside the purview of RCTs.

2.0 typology of RCTs focused on *how* and *why* different interventions work, to the 1.0 typology focused on *whether* interventions work. In a similar vein, Duflo (2017) remarks that (development) economists need to "adopt the mindset of a plumber," by focusing their research on how to improve last-mile service delivery questions. Intervention studies[7] for behavior change could be smaller, faster, and less expensive than multi-year health studies. They could iterate through trial and error, as Pritchett et al. (2013) recommend in their description of "learning in development projects." In many cases, it would make sense for such trials to be randomized, although what mattered most would be the careful recording of lessons learned from implementation. It is not without irony that we recommend WASH researchers adopt the same plumber mindset when designing intervention studies.

To make this recommendation practical, let us consider the sanitation Sustainable Development Goals of ending open defecation and increasing coverage of safely managed sanitation. Hence one recommended focus for these intervention studies would be reducing open defecation behavior, in places where doing so has proved difficult such as rural India (Rosenboom and Ban 2017). A notable effort towards this goal has been a series of studies organized by 3ie, the Gates Foundation, and r.i.c.e., intended to test strategies for promoting latrine use in rural India. The processes behind these studies and the results of the studies paint a less-than-rosy picture of how even the best designed "2.0" RCTs, focused on well-defined and policy-informed research questions, can struggle to produce policy-relevant evidence. After an iterative process of formative research, some promising pilot studies were selected for full intervention studies. The four full intervention studies were randomized. Even this well-crafted process, however, has been lengthy and its results difficult to interpret. Although the project started in 2015 as an effort to inform India's Swachh Bharat Mission (SBM), its results became available in the second half of 2019 and therefore have little chance of informing the Mission before its conclusion in October 2019. Furthermore, the results of the four studies (Friedrich et al. (2019), Chauhan et al. (2019), Visawanathan et al. (2019), and Caruso et al. (2019)) remain difficult to interpret. Across all four studies significant increases in latrine use (and reductions in open defecation) have been observed in both treatment and control clusters, presumably because of the high intensity of the SBM country-wide program, and/or because of the increased courtesy bias spurred by the heavy exposure to messages promoting latrine use.

[7] We use the term "intervention study" to mean a study in which an intervention is designed or modified specifically in order to learn about its effectiveness. Not all intervention studies are RCTs. Intervention studies, in our interpretation, differ from observational studies which use variation that occurs naturally in order to learn about effectiveness of interventions.

Another recommended focus for these intervention studies would be on safely managed sanitation,[8] particularly in urban areas where multiple behaviors by multiple agents are required to minimize the release of untreated excreta in the proximate environment. Houde et al. (2017), for example, study a supply-side intervention: increasing competition among vacuum-truck operators to reduce the price of safely emptied septic tanks. It should be noted that the cost of carrying out either of these intervention studies is between 10 to 30 times lower than the cost of the WASH-B or SHINE studies.

6.4 Conclusion: Good Use of Good Evidence Is the Only Standard

In WASH for development, there is no gold standard other than careful thoughtful research. In practice, this requires the collaboration of researchers from different backgrounds, with different expertise. While word counts in empirical social science journals are typically higher than the Lancet Global Health's 4500 we do encourage, we hope this does not remain a significant barrier to collaboration. It also requires judgment about better and worse evidence of each type, and the contexts to which evidence is likely to apply. Deaton and Cartwright would not be surprised by the experience of the WASH sector: "any special status for RCTs is unwarranted. Which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known. When little prior knowledge is available, no method is likely to yield well-supported conclusions."

To some researchers, the solution is more RCTs. According to their published protocols, fieldwork towards the SHINE and WASH-B trials started in 2011 and 2012, respectively. Brainstorming, planning, securing funding, and hiring staff surely started years before. The results were released for public discussion in 2018. These studies were impressive, complex achievements. Fulfilling them spanned most of a decade. For better or worse, these studies happened at a time when sanitation policy and practice were changing quickly in the developing world. According to UNICEF-WHO statistics, the fraction of people worldwide with safely managed sanitation is increasing by almost a percentage point a year. Open defecation is targeted to be eliminated by 2030 in the Sustainable Development

[8] The studies in this 3ie Thematic Window, completed by August 2019, show another practical challenge of using RCTs to learn about effectiveness in the context of a very large and politically important program. The government implementation of the Swachh Bharat Mission took such an intensity that reductions in open defecation were observed in both treatment and control populations (plausibly because the effectiveness of the government program was greater than that of the specific targeted interventions in the RCT and/or because the self-reported open defecation was highly susceptible to bias because everybody knows they have to report using latrines).

Goals. Another round of three trials as complex as SHINE and WASH-B might take most of the decade that remains before that deadline.

Fortunately, the WASH sector has much "prior knowledge." Although the three recent RCTs have attracted much discussion, almost nobody who previously believed that exposure to fecal pathogens harms children's development has changed their mind. Perhaps these recent studies tell us not to invest money in upgrading unimproved latrines into improved latrines; perhaps they highlight the importance of population density, of many behavior change visits, or of externalities beyond a single child's household. Nobody yet knows, and the RCTs did not settle the questions. This conversation will continue, and will continue to draw on diverse sources of evidence, as it should.

7

Microfinance RCTs in Development

Miracle or Mirage?

Florent Bédécarrats, Isabelle Guérin, and François Roubaud

7.1 Introduction

The rise of microcredit and the spread of randomized control trials (RCTs) marked two major milestones in development policies for poverty reduction in recent decades (Cling, Razafindrakoto, and Roubaud, 2003). The first took off in the 1990s and reached its zenith in the early 2000s with the launch of the UN International Year of Microcredit (2005) and the award in 2006 of the Nobel Peace Prize to Mohammad Yunus and to the Grameen Bank that he founded. The second became a thundering success a decade later, with RCTs acclaimed as the gold standard method for impact evaluations and, in 2019, another Nobel Prize (in economics) awarded to Esther Duflo, Abijit Barnerjee, and Michael Kremer, leading members of the RCT movement. These two developments are actually closely interlinked: microcredit was one of the flagship topics, an emblematic subject, to be evaluated by random experiments in development.

This chapter presents a detailed examination of RCTs on microcredit in development drawing on a wide range of analytical tools used in statistics, political economy, sociology, and development anthropology. Its main focus is the special issue (hereafter, the Special Issue) published in 2015 in a major economics journal—the *American Economic Journal: Applied Economics* (AEJ:AE). This Special Issue brings together six RCTs on microcredit, and the papers are prefaced by a general introduction (hereafter, the General Introduction) drawing broad conclusions. The Special Issue has had a great impact in both academic and professional circles, and tends to be seen as the definitive conclusion on the (limited) impacts of microcredit. But is it really?

We discuss this Special Issue from two angles: (1) top-down with a test on a specific case (microcredit) of the general criticisms made of RCTs, especially those developed by the authors in a previous article (Bédécarrats, Guérin, and Roubaud 2019); and (2) bottom-up with a study of the implementation of RCTs on the ground. We take as a starting point our replication of one of the six RCTs discussed in the Special Issue: the RCT conducted in rural Morocco (Bédécarrats

Florent Bédécarrats, Isabelle Guérin, and François Roubaud, *Microfinance RCTs: Miracle or Mirage?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: .1093/oso/9780198865360.003.0009

et al. 2019a, 2019b), which plays a central role in the Special Issue's "economy." We then expand the focus from the Moroccan case to take a more general angle by identifying the invariants that hold in other RCTs and ascertaining each RCT's particularities. More broadly, the main question we ask in this chapter is, "What lessons can be learned from RCTs on microcredit and how can their worldwide success be explained when they are not robust?"

The rest of this chapter is organized as follows. After summarizing the main features of the six experiments, the second part presents their main results and situates the Special Issue in the general context of the weight and role of microcredit in the RCT industry. The third part takes a comparative view to identify the main technical criticisms that can be made of this corpus of experiments, in terms of both their internal and external validity, as well as the ethical concerns raised. Moving beyond the method and the quantitative results, the fourth part analyzes the interpretations proposed by the authors (particularly in the General Introduction), and their underlying theory of change. In conclusion, we propose an interpretation of the hiatus outlined above—a far-reaching success despite major shortcomings—and we draw more general lessons from our work.

7.2 The RCT on Microcredit: A Sinking Flagship Product?

Microcredit is one of the main services provided by microfinance, one of the sectors the most frequently evaluated by RCTs. An illustration of this importance can be found on the RCT online repository managed by J-PAL (a global research centre promoting this method for poverty reduction and the leading provider and promoter of RCTs). In 2010, this repository displayed 233 RCTs, of which 32 percent were labelled as "microfinance" (Bédécarrats, 2012). JPAL since then reorganized its evaluation labelling with broader categories, and it currently posts 287 "finance" RCTs out of its 978 RCTs.[1] Finance is J-PAL's foremost sector of interest, ahead of Education (233), and Political Economy and Governance (216). Although microfinance is just a subset of "Finance" RCTs, J-PAL is a major provider of impact evaluations on the subject. The mid-2000s saw a boom in the number of RCTs on microfinance and the RCT industry as a whole (Bédécarrats, Guérin, and Roubaud 2019; and Ravallion, Chapter 1, this volume). Since then, the number of microfinance RCTs has dropped sharply while RCTs in general have continued to grow (Figure 7.1). There is no easy way to count the number of RCTs conducted worldwide. Our estimates are based on 3IE's online impact evaluation repository rounded out by Bédécarrats (2012)

[1] Source: The Abdul Lateef Jameel Poverty Action Lab website: www.povertyactionlab.org/evaluations, visited on 10/13/2019.

and J-PAL's online evaluation repository.[2] Figure 7.1a illustrates that the impact of microfinance has long been a disputed issue, generating numerous non-experimental impact evaluations. Despite the fact that experimental methods provide theoretically stronger quantitative empirical evidence, non-experimental studies furnish a wealth of relevant evidence. There has also been a sharp increase in experimental evaluations, coinciding with a sharp decrease in non-experimental evaluations, although these trends might be marginally exaggerated by omissions of the most recent studies in the registries we used. Figure 7.1b also shows that microfinance was a prominent theme for the randomista[3] movement up to 2013,

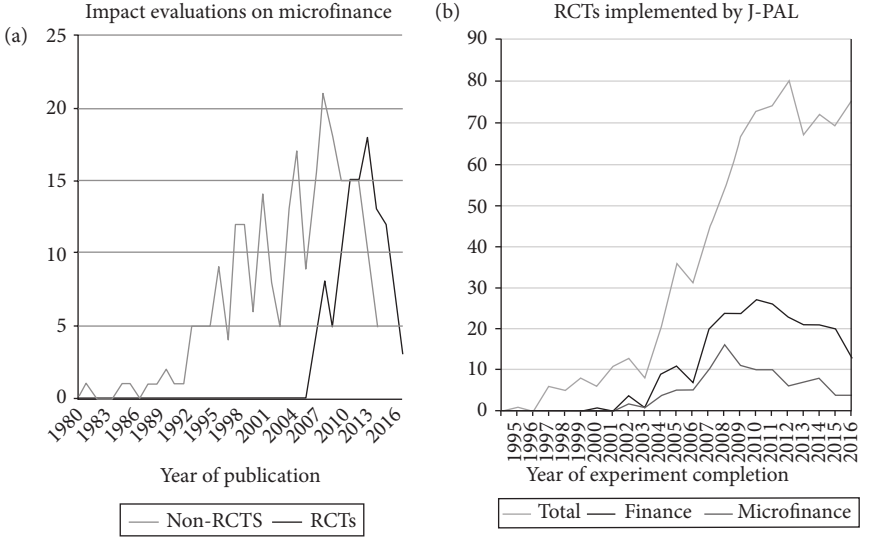

Figure 7.1 RCTs on microfinance

Source: Authors, based on: 3IE evaluation repository (2019), J-PAL evaluation repository (2019) and Bédécarrats (2012) for Panel (a); and J-PAL online evaluation repository (2019) for Panel (b).

[2] 3IE's online impact evaluation repository forms the main catalogue of results of impact evaluations on development interventions (https://www.3ieimpact.org/evidence-hub/impact-evaluation-repository, last accessed for the authors' update on October 13, 2019). 3IE tends to underreport non-experimental evaluations and its inventory work appears to have dropped off in recent years, as references decrease from 2015 onwards. We have rounded out 3IE's data with the impact evaluations listed in Bédécarrats (2012) and references included in J-PAL's evaluation repository. References have been matched to avoid double counting the same evaluations. Figure 7.1b is based on the references listed in J-PAL's online evaluation repository (https://www.povertyactionlab.org/evaluations, accessed for the last update on October 18, 2019). "Finance" in the key is the label assigned by J-PAL to the registered evaluation. The authors assigned the "Microfinance" label after reviewing the summaries of all the evaluations registered as "Finance" on J-PAL's website. The dates in Figure 7.1b correspond to the year in which the experiment was completed, while the dates in Figure 7.1a stand for the year in which the experiment results were published.

[3] We call randomistas those RCT proponents who are convinced that RCTs are the only way to rigorously assess impact in evaluation, and that they are superior to other methodologies in all cases.

but that interest has since waned. The fall in the second half of the 2010s following the peak in the first half is intriguing: is it due to a trend shift or is it because there is not much left to say about this overstudied issue? This is one point we will address in the following.

It was at the height of RCTs in microfinance that a 2015 special issue was published in the *American Economic Journal: Applied Economics* (AEJ:AE) featuring six RCTs on microcredit (Banerjee, Karlan, and Zinman 2015). This special issue is seen by leading RCT movement figures as the decisive contribution to settle a long-standing debate on the subject (Ogden 2017), both in academia and among donors and policy-makers. It quickly attracted massive coverage, as seen from the 3,607 citations of its articles in other scientific publications.[4] In a move to promote its use to inform policy-making, J-PAL and IPA published a policy briefcase that took stock of the special issue and drew general conclusions for microcredit worldwide (J-PAL and IPA Policy Bulletin 2015). Some researchers even mused that it might be the "last word on microcredit" (Sandefur 2015).

Looking more carefully at the academic impact of the AEJ:AE Special Issue, the result is impressive. Google Scholar (accessed 13/10/2019) lists the General Introduction alone as having been cited 527 times. A great performance, although way behind the paper by Banerjee et al. (2015b) on the Spandana microcredit programme in India (1,813 citations). The other five papers have also performed very well: 320 citations for Angelucci, Karlan, and Zinman (2015) on Compartamos Banco in Mexico, 298 for Crépon et al. (2015) on Al Amana in rural Morocco, 225 for Attanasio et al. (2015) on Mongolia, 214 for Augsburg et al. (2015) on Bosnia, and 210 for Tarozzi, Desai, and Johnson (2015) on Ethiopia. By way of comparison, the count for Pitt and Khandker (1998), quoted by Roodman and Morduch (2014) as the all-time most cited empirical article on an individual microcredit project, stands at 1,956 citations more than twenty years after its publication.

In addition to direct citations, the Special Issue's impact is cascaded through quotations of citations (like any article), but also through systematic reviews or meta-analyses, which build mostly on the Special Issue as their main body of evidence (Brody et al. 2015; Buera, Kaboski, and Shin 2015; Chernozhukov et al. 2018; Demirguc-Kunt, Klapper, and Singer 2017; Meager 2019). Special mention can be made of the article published in the prestigious *Science* review in 2015 (Banerjee et al. (2015a), cited 484 times).[5] This article extensively discusses the Special Issue, highlighting the comparative merits of a different approach ("graduation" programmes).

[4] Source: Google Scholar citation indexes on the articles featured in this special issue, see corresponding webpage, visited on 10/13/2019.

[5] This is not the first time that *Science* has opened its columns to RCTs on microcredit (Karlan and Zinman 2011).

Lastly, the results of the Special Issue have circulated widely beyond academic circles to the world of microfinance practitioners (J-PAL and IPA Policy Bulletin 2015). CGAP, which plays a leading role in disseminating good practices in the microcredit sector, commented on it even before its release (Cull, Ehrbeck, and Holle 2014). For many practitioners (whom one of us meets regularly in conferences and in the field), the results of the Special Issue are now conventional wisdom.

Ultimately, whether judged on the basis of the number of RCTs conducted or the dissemination of results, microfinance, and microcredit impact evaluations in particular, do appear to be the flagship products of the franchise created by the randomistas based on the RCT method, and the Special Issue the outstanding prototype for this movement.

7.2.1 A Focus on the Design of the AEJ:AE Special Issue

The Special Issue features six articles on six microcredit RCTs conducted by six affiliated J-PAL teams in six different countries (Bosnia and Herzegovina, Ethiopia, India, Mexico, Mongolia, and Morocco) at around about the same time (from 2006 to 2012). It is preceded by a General Introduction that draws general lessons from this collective experience. The Special Issue draws its strength from a downstream harmonization process organized by the journal in preparation for its publication.[6] A common analysis plan was drawn up to facilitate comparisons. As far as possible, the impact of microcredit was estimated using the same econometric methodology for a set of common outcomes, themselves calculated the same way. This was the first time that such a pooling effort had been made on this scale. It represents a decisive advantage when it comes to generalization.

Not only does the Special Issue appear decisive in terms of results, but it also marks a 'good practices' shift by RCT proponents. Hence the issue seeks to address a number of limitations. For the first time, the issue as a whole, and the General Introduction in particular, provide elements of response to five types of recurrent criticisms of the pro-RCT movement (Bédécarrats, Guérin, and Roubaud 2019): a theoretical model is developed in response to the agnostic empiricism criticism of RCTs; a cost-benefit analysis is proposed to answer the question of effectiveness, to move beyond mere causal impact; the issues of take-up rate, estimator accuracy and treatment heterogeneity are acknowledged and discussed; contextual diversity is addressed by a range of settings, products, and institutions covered by the six papers, enabling the Special Issue's editors to claim their sample is "*fairly*

[6] "Drawing lessons across the six studies has been greatly facilitated by the efforts of the six research teams and the editor, Esther Duflo, to make the papers readily comparable" (Banerjee, Karlan, and Zinman 2015: 2).

representative of the microcredit industry/movement worldwide" (Banerjee, Karlan, and Zinman 2015: 2); and, lastly, the Special Issue professes to make available the original databases in response to the complaint about replicability and in order to facilitate meta-analyses.

Let us briefly describe the six RCTs. Despite an upstream harmonization process (data processing and analysis), the experiments differ significantly in their protocols. The types of microcredit products, microfinance institutions (MFIs hereafter), unity of randomization procedures, and so on vary from one RCT to another. The authors interpret this diversity based on the assumption that the similarity of results across this wide range of environments is a guarantee of their robustness, and therefore evidences the generic properties of microcredit impacts; a way of addressing the recurrent criticism of RCTs as lacking external validity.

The General Introduction gives a detailed presentation of the main features of the six RCTs, summarized in Table 7.1. The MFIs vary in size, with some being commercial while others are not. We find all kinds of products: joint liability and individual loans, weekly, and monthly repayments, an annual interest rate varying from 12 percent to 110 percent (on average), and the (average) loan amount ranging from 6 percent to 118 percent of monthly income. Half of the microcredit programmes target women. In terms of geographic areas, one is exclusively urban (India), three are exclusively rural (Ethiopia, Mongolia, and Morocco) and the remaining two cover both types of area. One point of note is that, in all cases, the client eligibility criteria are ad hoc: they depend on both the internal rules of each MFI and on the parameters of each RCT. As a result, the target populations are highly specific (if not unique), undermining the possibilities for inference and extrapolation to larger populations; we will come back to this point in the third part.

7.2.2 A Focus on the AEJ:AE Special Issue: Main Results

The General Introduction draws seven major lessons from the exercise. In the first place, low take-up is a constant in all the studies except Bosnia, leading to the conclusion that microcredit cannot be the universal panacea for lifting the poor out of poverty. An unfortunate consequence of the low take-up is that it poses a problem of statistical power and a challenge for the RCT identification strategy. However, the General Introduction puts forward the Moroccan, Indian, and Mexican RCTs to provide new elements to address these shortcomings (take-up prediction and sampling strategy). Second, and tying in with the previous point, it is particularly difficult to predict the take-up rate, and no study has managed entirely satisfactorily to do so. Third, and probably the main conclusion, access to microcredit is not transformative either for

Table 7.1 Main characteristics of the six RCTs

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Interest rate (APR)	22%	12%	24%	110%	27%	14%
Liability	Individual	Group	Group	Group	Both	Group
Average loan/ household income	9%	118%	22%	6%	43%	21%
Sex of potential clients	Both	Both	Female	Female	Female	Both
Loan eligibility (among other)	Strong collateral, repayment capacity, creditworthiness . . .	Poverty status, business plan. . .	18–59 years old, proof of residence home ownership. . .	18–60 years old, valid ID card, proof of address. . ..	Assets < $869 Profit < $174/ month	18–70 years old, ID card, non-livestock agricultural activity. . .
Area coverage (urban/rural)	Both	Rural	Urban	Both	Rural	Rural
Area coverage (regions/cities)	14 (nationwide)	2 (Western)	1 (City)	4 (NC Sonora)	5 (North)	11 (nationwide)
Unit of randomization	Individual	Association	Neighborhood	Neighborhood & village	Village	Village
Final Sampling Unit	Risky and unreliable applicant . . .	Random households	Household with >=1 woman >=3 years in the area. . .	Has a business or would like one . . .	Interested in obtaining a loan . . .	Household deemed likely borrowers . . .
Sample size (endline)	995	6,263	6,862	16,560	964	5,551

Source: Authors, based on Banerjee, Karlan and Zinman (2015c) (Tables 1 and 2).

microenterprise performance or for household living conditions—including social well-being and women's empowerment—at least on average. The only robust finding for consumption is a decrease in "discretionary spending," defined by the authors as "temptation goods, recreation/entertainment/celebrations" (Banerjee, Karlan, and Zinman 2015: 13). Fourth, only firm investment is stimulated by microcredit, showing that it cultivates micro-entrepreneurs' intentions to develop their business. Fifth, other modest, albeit potentially important effects are pointed up: freedom of choice in particular. Sixth, although microcredit is not transformative, it does not have any catastrophic effects either, which places proponents and opponents of microcredit on a level pegging. Lastly, the seventh lesson relates to the presumption of heterogeneity of microcredit impact, which could be positive (even transformative) for some (the upper tier), and negative for others. This brings us back to the issue of statistical power, the sample sizes required to properly estimate impacts and the representativeness of the targeted populations. Table 7.2, based on the General Introduction and the J-PAL and IPA Policy Bulletin (2015), summarizes the results obtained by the six RTCs for the main outcomes monitored.

Table 7.2 Main results of the six RCTs

	Bosnia and Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Business ownership	Positive	n.s.	n.s.	n.s.	Positive	n.s.
Business revenue	n.s.	n.s.	n.s.	Positive	n.s.	Positive
Business assets	Positive	–	Positive	–	Positive	Positive
Business investment	n.s.	n.s.	Positive	Positive	–	Positive
Business profits	–	–	–	–	–	Positive
Household income	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Household consumption	n.s.	Negative	–	Negative	Positive	–.
Household consumption of temptation goods	Negative	–	Negative	Negative	n.s.	Negative
Social well-being	n.s.	n.s.	n.s.	Positive	–	n.s.
Women's empowerment	–	n.s.	–	Positive	–	–

Note: n.s. (not significant at 10 percent); —(no data).

Source: Authors, based on J-PAL and IPA Policy Bulletin (2015); Banerjee et al. (2015c).

In conclusion, the Special Issue is considered by many, starting with the authors themselves (Ogden 2017), as the most comprehensive summation on the impact of microcredit. Its general conclusions have scarcely been questioned since its publication in 2015 (for exceptions, see Dahal and Fiala 2020; Wydick 2016). In a way, it freezes the state of the art on the causal impacts of microcredit and its role for development and poverty eradication. For AEJ:AE's editors, and subsequent papers elaborating on the six RCTs, the Special Issue does even more than this. It is praised for pushing back the frontiers of scientific knowledge, both on microcredit and on the RCT method. Three papers, posterior to the Special Issue and directly following up on the same set of RCTs, are good illustrations of this. Meager's (2019) article, published again in AEJ:AE, confirms that it is still considered the must on microcredit. This article takes the six RCTs in the Special Issue (plus an RCT in Philippines; Karlan and Zinman 2011) to re-estimate the general impact on the main variables and answer the question of external validity using an innovative method (a Bayesian Hierarchical Analysis). Then there is Chernozhukov et al. (2018), who apply a double machine learning method to study heterogeneity in this data set. A third example is Banerjee, Duflo, and Kremer (2019), published as this chapter was being written. The paper draws on a third-round survey for the *Spandana* Indian RCT. While responding to some of the criticisms of RCTs (by addressing heterogeneous treatment, lengthening the time span and developing a theoretical model), the paper largely refers to and takes stock of the Special Issue, presented as the seat of knowledge on microcredit to date. This paper may not be the last in the series. In the same vein, Crépon et al. (2015) also announce in the conclusion to their paper a third-round survey for the Moroccan RCT to assess the long-term impact of microcredit.[7]

7.3 Validity and Scope of the Special Issue: A Critical Assessment

In the literature, RCTs are appraised from two main angles: external and internal validity. External validity is pivotal when it comes to scaling up, informing and designing public policies on a broader scale (national or regional) and to testing a theory. Internal validity is usually taken for granted with RCTs, and seen as their major strongpoint over other methods. While this property may be true in theory, implementation constraints in the field can call these ideal conditions into question, a point hitherto overlooked.

[7] "We are currently following up with the households, now that a much longer time period has elapsed, to check if the investment in business assets paid off in the longer run" (Crépon et al. 2015: 148).

7.3.1 Internal Validity

Assessing the internal validity of RCTs calls for a probe into the making, and tinkering, of RCTs in the field. We performed this demanding exercise on the Moroccan study (Crépon et al. 2015). We present below the main results of the two companion papers we produced from this review (Bédécarrats et al. 2019a, 2019b).

The Emblematic Case of the Moroccan RCT

From 2006 to 2010, a research team from J-PAL conducted an RCT in rural Morocco to measure the impact of microcredit provided by Al Amana, then the Moroccan market's leading MFI, in the midst of a phase of expansion.

We replicated Crépon et al.'s paper and identified a number of issues that challenge their conclusions (Bédécarrats et al. 2019a). We argue that they used inconsistent trimming procedures and thresholds, and that their results depend heavily on how their data was trimmed. Crépon et al. (2015) reported a balanced sample at baseline after removing extreme values on 24 variables over 459 observations (10.3 percent of the sample). At endline, however, they trimmed 27 observations (0.5 percent of the sample) differently by removing them entirely. Moving the endline trimming threshold by just 0.2 percent (removing a dozen observations more or less) produces radically different results in terms of sales, expenses, investment, and profits. No other trimming threshold would have produced results consistent with their published findings and no other paper in the same special issue used a similar trimming method or threshold.

We found substantial and significant imbalances in the baseline for a number of important variables, including the RCT's outcome variables. Possibly in relation to this, we estimated implausible "treatment effects" on certain variables, e.g. on the household head, gender, and spoken language. We documented numerous coding errors. For instance, the appraisal of agricultural assets at endline omitted two types of assets (tractors and reapers), which happen to be the most valuable assets owned by surveyed households. Inclusion of tractors and reapers in asset appraisal increases the sample's average value of agricultural assets per household by 470 percent (from 1377 Moroccan Dirham to 5111 Moroccan Dirham). The identified coding errors altered some 80 percent of the observations.

Inconsistencies in credit measures warrant particular attention, as they are essential to characterize the treatment evaluated by this experiment. Crépon et al. (2015) append administrative data to the survey data, reporting the former's given microcredit take-up of 17 percent rather than the latter's 11 percent. They contend that the Moroccan population underreport borrowing because of religious shame. However, we argue that this is implausible as the inconsistencies between sources go way beyond differences in averages. A total of 195 of the 435 reported clients said they had never borrowed from the MFI. However, a "credit

shame" explanation for these households would imply a "credit pride" explanation for the 152 households that reported having a loan from the MFI even though they did not appear on its registers. According to the survey data collected on the panel sample, access to credit remained stable in the treatment group between baseline and endline, while it was decreasing in the control group (there was a major crisis in Moroccan microfinance from 2008 to 2010). Our results challenge the very meaning of this RCT: what was tested appears to have been not the impact of the introduction of microcredit in "virgin" areas, but rather the replacement of other formal sources with one microcredit source in the treatment group and credit rationing in the control group.

We also found sampling errors. For example, the sex and age composition for 20 percent of the households interviewed at baseline and reportedly re-interviewed at endline differs to such an extent that it is implausible that the same units were re-interviewed in these cases. In addition, we found that Crépon et al.'s sample characteristics differed in substantial ways from the population's characteristics. The average number of household members grew from 5.17 to 6.13 between the baseline and endline surveys. The national census, however, reported that Moroccan rural households had an average of 6.03 members in 2004 and 5.35 members in 2014. Such discrepancies raise questions about the sample's representativeness, and hence undermine the external validity of this study.

The authors produced a reply to our replication, entitled *Rejoinder*, rejecting most of the errors we documented (Crépon et al. 2019). They referred to our analysis, but they do not appear to have replicated or closely analysed its statistical content and we argue that their rejoinder thereby contains numerous factual errors and omissions. We published a review of their main arguments in response to our replication (Bédécarrats et al. 2019c). We found that all the coding, measurement, and sampling errors documented in our replication still hold.

Distortion of the Protocol: Product and Sampling Tweaking

Our second paper sought to explain how such inconsistencies could occur, using a qualitative field study specifically designed to round out the RCT (Morvant-Roux et al. 2014) and various data and documents, public and internal from the RCT's key stakeholders (Bédécarrats et al. 2019b). The paper describes the entire study production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of results. Far from ideal laboratory conditions,[8] the analysis of the randomized protocol's implementation on the ground by the different players (each with their own motivations and constraints) finds a number of discrepancies compared with the theoretical protocol reported in the published article.

[8] Field experiments such as RCTs are designed precisely to get out of the artificial world of laboratories. But too often randomists think that the protocol can be applied as it is, as in the laboratory, which is not the case.

A major concern during the study was take-up, much lower than initially expected, which prompted a number of corrective measures. The first tweak was to modify the intervention (microcredit supply) by launching further information campaigns, introducing one-off bonuses for agents, and withdrawing the minimum quota for women. Take-up became an "obsession" for both research team and loan officers, who used the term themselves and went to great lengths to convince villagers to take out microcredit. Strategies included pushing back the usual village borders in the hope of finding more clients.[9] When these measures proved insufficient, the team tweaked the sampling method (modification of prediction models, and addition of new households at endline, with a supposedly higher propensity to borrow). Villages with zero take-up were dropped.

Poor Data Quality and Measurement Errors

Data collection and entry were subcontracted to a consultancy firm specialized in engineering, but with no experience of statistical surveys. For the purpose of monitoring the RCT's design and implementation, the RCT's funder (AFD) appointed a team of economists and specialists in household surveys. The team reporting back on its field missions found serious data collection dysfunctions at an early stage. These included translation problems because interviewers did not speak Berber, a language spoken by a large part of the target population. Interviewers therefore made extensive use of impromptu translators, including local leaders, raising comprehension and response bias problems (social desirability and mistrust of government).

Another concern was the number of respondents in households and extended families, which again appeared to be improvised depending on the presence and availability of people and their ability to understand each other and the interviewers. These observations probably explain in part the above-mentioned significant discrepancies between baseline and endline. However, the size of the gap suggests another explanation: some households may not have been the same, as confirmed by our replication. Absence of a precise address calls for precise tracking techniques, which may have been overlooked. Lacking time and supervision, some interviewers may simply have interviewed households available at the time of their visit. AFD's team made recommendations to improve the quality of the data collected, expressing concerns about the potential repercussions of these shortcomings on the experiment's results. They also raised the data entry issues. Although the J-PAL team responded, challenging the gravity of the problems and contending that they did not call into question the internal validity of the experiment, the next steering committee meeting decided that all questionnaires already

[9] Changing the product for the sake of the RCT is also an external validity issue (as experimental conditions are not in line with how it functions in the "real world" (Peters, Langbein, and Roberts 2018).

entered were to be sent to the French National Statistical Office (INSEE) in Paris to be re-entered.

These different issues were omitted from the published article and point to shortcomings in the preparation, implementation, and follow-up of fieldwork.

Beyond the Moroccan RCT: A General Assessment

It is not feasible to analyze the other five RCTs in the Special Issue in such detail, both for reasons of time and because the necessary raw data are available only for two of them (Table 7.4). We therefore perform a partial exercise, namely a critical reading on the usual review summary terms, i.e. based on the published articles. Table 7.3 summarizes the internal validity problems as they can be assessed from the information available to us. Hardly any of these problems are addressed by the Special Issue, and even less so by the General Introduction. We discuss here the sampling error and measurement error issues in turn.

With regard to sampling, note that the papers generally do not provide the basic elements to be able to accurately describe and qualify the adopted sampling designs and selection plans (the standards for such descriptions are provided, for instance, in Statistics Canada 2010 and Ardilly and Tillé 2006). The authors focus their analyses on randomization and causal inference issues. First, the reference population is never clearly established. In most cases, it corresponds to eligible clients in the MFI's expansion areas, although it is not known how the latter are defined. This has unfortunate repercussions on the external validity of the RCTs (see Section 7.3.2). Second, the adopted sampling plans fall into the general category of multi-stage stratified random sampling, with the exception of the RCTs in Bosnia and Herzegovina and Mongolia. Neither of these cases is randomly sampled: in Mongolia, the first 30 poor women in each selected village to state an interest in obtaining a loan were selected; in Bosnia, loan officers were asked to select potential clients who were not deemed eligible by the current MFI's standards. In all cases, these complex sampling designs, to use statistical terminology, either do not enable the confidence intervals associated with the estimated impact to be computed (the above-mentioned two cases) or would call for particularly complex variance estimation calculations, which are not performed (except for estimating cluster-robust standard errors). The direct consequence of this gap is that the confidence intervals are probably underestimated and the impacts deemed significant, already small in number, should not be statistically different from zero.

Moreover, four of the six RCTs deviated from the experimental method's canonical protocol: random selection of a treatment and control group, a pre-treatment baseline survey (BL) and then panel monitoring based on a post-treatment endline survey (EL). In the Ethiopian case, the baseline and endline surveys were not on panels, but cross-sections (i.e. different individuals were surveyed). This makes it impossible to identify potential imbalances at baseline

Table 7.3 Internal validity of the six RCTs

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Population of interest	**Potential clients** initially rejected by the MFI as uncreditworthy	**Rural households in two ad-hoc areas**	**Likely borrowers** (women living in slums for more than three years with valid ID) in MFI expansion areas in Hyderabad	**Potential clients** (women owning or planning to create a business or intending to borrow) in MFI expansion areas in Central Sonora, Mexico	**Poor women:** (assets <$869 & profit<$174/ month) **Signed up to get a loan**	**High borrowing propensity households in rural MFI extension areas**
Sample design, randomization						
Sample design	Purposive individual sample	- Stratified (2 "zones") - 3 degrees (Admin units/Village/HH)	2 degrees (slums/HH)	2 degrees (village/HH)	- Stratified (5 provinces), - 2 degrees (village/HH) Northern Mongolia	2 degrees (village/HH)
Info on area selection #of areas (T, C)	Not applicable	Yes 353 villages	Yes 104 (52, 52)	Yes 238 (120, 118)	No 25 (15, 10)	Yes 162 (81, 81)
Discarded areas	Not applicable	No	Yes (16 slums)	Yes (12 areas)	No	Yes (not specified)
Info on selection of individuals	Yes, not random	Yes, random	Yes, random	Yes, random	Yes, not random (1st 30 to sign up)	Yes, random
Info on randomization (T vs C)	Yes (individual level)	Yes (village level)	Yes (slum level)	Yes (area level)	Yes (village level)	Yes (village, level)
Sample size (full; control)	BL (1,196; 568) EL (994; 444)	BL (6,412; n.a.) EL (6,263; n.a.)	BL (2,800; 1,220) EL1 (6,863; 3,264) EL2 (6,142; 2,943)	BL (6,786; n.a.) EL (16,560; 8,298)	BL (710, 299) EL (610, 260)	BL (4,465; 2,266) EL (5,551; 2,810)

(Continued)

Table 7.3 (*Continued*)

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Attrition rate (BL- > EL): %total, %control	Panel (17%; 22%)	No panel 2 cross-sections	BL- > EL: no panel EL1- > EL2 (11%; 10%)	Panel (37%; n.a.)	Panel (16%; 15%)	Panel (8%; 7%)
Respect of experimental design	Yes	No 22% areas misallocated (12% T not treated), 23% C treated)	No (16 areas dropped; BL unreliable)	No (BL aborted)	Yes	No (new HHs added at EL)
Balance tests at baseline						
Population included	Panel households only	Panel households only	All BL households	Panel households only	Panel households only	All BL households
Tested variables	27	35	33	14	48	43
Include main study outcomes	Yes	Yes	Yes	No	Yes	No
Reported significant imbalances	No	No	No	Yes	Yes	Yes
Trimming	Results with and without 1% trimming for robustness checks	Results with and without trimming 8 obs for robustness checks	No	No	No	BL: trim highest values for 10.3% of obs. EL: trim 0.5% of obs.
Data quality (discussion in paper)	No	Yes, marginal (measurement errors)	Yes, marginal (possible recall errors)	Yes, marginal (missing outcome variables at EL)	No	No (except take-up admin vs survey)

Source: Authors based on AEJ:AE (2015). Notes: HH: households, BL: baseline survey, EL: endline survey, T vs C: treatment group versus control group, obs.: observations.

Table 7.4 External validity, acknowledged caveats, and ethical concerns

	Bosnia & Herzegovina	**Ethiopia**	**India**	**Mexico**	**Mongolia**	**Morocco**
Population of interest	MFI expansion	MFIs expansion	MFI expansion (partial)	MFI expansion	MFI expansion	MFI expansion (partial)
Extrapolation to any superpopulation?	No	No	No	No	No	No
Potential threats discussion (in paper)**?**						
Hawthorne or John Henry	Yes	No	No	No	No	No
General equilibrium	No	No	No	No	No	Yes
Comparison with NSO data?	Yes	No	No	No	Yes	No
Other surveys/ methods implemented?	No	No	No	No	Village surveys, qualitative interviews	No
If yes used?	–	–	–	–	No	–
Explicit caveats acknowledged?	**Yes** 1- No external validity 2- Underpower 3- Potential H&JH effects	**Yes** 1- No external validity 2- Underpower 3- No panel= Imbalance at BL, selective attrition, heterogeneous effect 4- No respect of experimental design 5- No Consumption 6- Measurement errors	**Yes** 1- Underpower 2- Non-Representative BL 3- Selective Attrition and migration 4- Contamination 5- ITT representative of "likely borrowers" only	**Yes** 1- No external validity 2- Data quality 3- No BL 4- Heterogeneous treatment periods	**Yes** 1- No External Validity 2- Underpower 3- Presence of other MFIs 4- Attrition (possible imbalance) 5- Not robust at MHT	**Yes** 1- Small significant imbalances at baseline

(Continued)

Table 7.4 (*Continued*)

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Ethical concerns discussion						
Informed consent for experiment	No	No	No	No	No	No
Risk analysis and monitoring	No	No	No	No	No	No
Equipoise	No	No	No	No	No	No
Reproducibility						
Data available	Raw data	No	Aggregated data	Aggregated data	Raw data	Raw data
Detailed code available	Yes	No	Partially	Partially	Yes	Yes
Survey questionnaire available on AEJ	Yes	No	No	No	No	Yes

Source: Authors, based on papers published in *American Economic Journal: Applied Economics* volume 7, no. 1. Note: MFI: microfinance institution; NSO: National Statistical Office. ITT: Intention to treat; BL: baseline survey; EL: endline survey; H&JH: Hawthorne and John Henry, MHT: Multiple Hypothesis Test.

for the population for which impact is estimated at endline. In the Mexican, Moroccan, and Indian cases, the field surveys could not be conducted as initially planned and threats to the experiment's success led to the initial protocol being readjusted along the way. In India, the baseline did not constitute a base panel for either of the two subsequent endlines,[10] raising the same problems as in the above-mentioned case of Ethiopia. In Mexico, the baseline was aborted due to the poor quality of the data collected: 73 percent of the baseline households were not revisited at endline and 89 percent of the endline sample had not been surveyed at baseline, so the majority of the households surveyed at endline were added at this stage. A similar strategy was adopted in Morocco. Low take-up by households identified as potential borrowers meant that new households were selected at endline that represented 26 percent of the endline sample. If we also take into account the attrition rates (available only for the panel protocols) ranging from 8 percent (Morocco) to 37 percent (Mexico), it is clear that none of the RCTs was conducted in keeping with the standards (non-random sampling of targeted households in Bosnia and Mongolia, and non- or failed-panels for the other four due to data collection issues or low take-up).

However, it is fundamentally important to verify sample balance at baseline. The studies vary a great deal in terms of the variables tested. Some tested surprisingly few variables compared with the wide range of data collected (Mexico). Others tested many more, but all differ as to which variables were tested. In some cases, most of the variables include at least some of the outcomes for which impact was measured at endline. In the case of Morocco however, the balance tests were applied only to specific subsets of the outcome variables evaluated at endline (e.g. sales for crop farming households, or livestock breeding households, instead of overall sales reported at endline). In our replication, we found large, significant imbalances in these outcomes. Households in the treatment group made 22 percent less sales and profits from self-employment than households in the control group (significant at the 5 percent level). They also invested 61 percent more (significant at the 5 percent level). In addition, there are imbalances at baseline with respect to a number of important variables, such as the surface area of owned land, access to basic services and women's empowerment. In addition to the variables tested, the calculation basis is also important. For example, the Mexican study limited its balance tests to 1823 households surveyed at both baseline and endline. If we extend the same tests to all the households surveyed at baseline (6786), as is the case with India and Morocco, we find significant differences in household income per adult in the previous month, especially those in an informal group.[11]

[10] Banerjee et al. (2015b) conducted a first endline survey in 2007 and 2008. They re-interviewed the households of the first endline in a second endline survey in 2009 and 2010.

[11] Computations are available from the authors on request.

Even if baseline differences between treatment and control groups are not statistically significant, they can be very large. In Mongolia and Ethiopia, baseline balance tests found average differences often over 10 percent (and up to 50 percent), but not significant (not surprisingly given the small sample sizes). They are systematically interpreted for what appears to be convenience's sake (absence of imbalances and therefore success of the randomization process), while the opposite explanation is often given for the results: where coefficients are non-significant due to underpower, they are construed as being "economically meaningful."

None of the papers discusses measurement error issues in depth. However, the literature emphasizes how difficult it is to obtain reliable measurements of many of the outcomes analysed, especially household consumption and microenterprise and agricultural production (Deaton 1997; Grosh and Glewwe 2000). Measurement errors merely get a mention in a footnote on potential memory bias in the Indian case, and in a discussion on under-reporting of borrowing in the Moroccan case, said to explain the differences between the administrative data and the surveys on this subject. Only the Ethiopian RCT reports major data quality concerns, and explicitly acknowledges that this issue affects internal validity. The Mexican RCT specifies that the baseline had to be interrupted and that its data could not be used because they were unreliable, without providing any details or indicating how more reliable data could have been collected at endline. Unfortunately, it is impossible to discuss data quality further from the articles alone. However, a detailed analysis of data consistency and the recoding conducted by researchers in the Moroccan case (Bédécarrats et al. 2019a) shows that this problem altered the results. There is evidence to suggest that similar problems may exist in other cases. For instance, a preliminary analysis of the Mexican data finds that the age ranges do not match between surveys for 231 (12.7 percent) of the 1823 women interviewed in principle at both baseline and endline.[12]

7.3.2 External Validity

The question of RCTs' external validity is the most discussed in the literature. External validity is a key issue, especially since, in contrast to a lot of observational data, RCTs are conducted on a small scale and in non-representative locations as seen above. External validity is also at risk when sampling is selective, that is when a study focuses on specific sites and population categories. Then there is the implementers' bias: for instance, where the results obtained by an NGO do not replicate when the same intervention is delivered on a larger scale by a government (Bold et al. 2013; Vivalt 2017). However, the issue of the external validity of RCTs is rarely

[12] Banerjee et al. (2015b).

given serious consideration by the randomistas. Peters, Langbein, and Roberts (2018) conduct a systematic review of all (54) RCTs published in leading economic journals from 2009 to 2014 to assess the main threats to external validity (Hawthorn/Henry effects,[13] general equilibrium effects, specific sample problems, and special care in treatment provision). Based on a set of objective indicators, albeit with lenient criteria, the paper finds that the majority of published RCTs do not discuss these hazards and many do not provide the necessary information to assess potential problems.

External validity also has to do with the relevance of the selected results. The focus on an "average" impact and problems capturing the heterogeneity of impacts and their distribution form a major obstacle to the relevance of results (Ravallion 2009a; Stern et al. 2012; Vivalt forthcoming). The restriction to a short-term impact (for reasons of cost and attrition) often means that mid-point indicators are studied, which can be very different from final outcomes (Boone, Eble, and Elbourne 2013), if not vice versa, since many project trajectories are not linear (Labrousse 2010; Woolcock 2013). Knock-on and general equilibrium effects are overlooked, albeit partially in the Moroccan RCT, despite there being any number of them (Acemoglu 2010; Deaton and Cartwright 2018; Ravallion 2009a). The same holds true for the political consideration involved in programme replication, despite its being a key consideration for scale-up (Acemoglu 2010; Bold et al. 2013; Pritchett and Sandefur 2013b). Last but not least, the *reasons* for the impact are disregarded: RCTs might be able to measure and test some intervention impacts and aspects, but they cannot analyze either their *mechanisms* or their underlying *processes*. Notwithstanding the method's limitations, the absence of theory prevents any form of understanding of the processes of change. Overcoming this limitation of the probabilistic theory of causality would call for a "causal model" (Cartwright 2010), a coherent theory of change (Woolcock 2013), a structural approach (Acemoglu 2010), and evaluation of the intervention in context (Pritchett and Sandefur 2015; Ravallion 2009a).

Table 7.4 summarizes the problems of external validity as they can be assessed from the information available to us. The usual RCT shortcomings hold here.

First, the sampling is selective: the experiment's selection criteria are ad hoc since they were conducted in MFI extension zones. As Wydick (2016) shows, the constraint of randomization (identifying virgin areas or populations) forced randomistas to choose "marginal" areas and populations previously neglected by MFIs and therefore highly specific in relation to the "normal" market. The unsuccessful bids by the Moroccan, Mexican, and Indian studies to identify likely borrowers

[13] These are behavioral biases induced by the experiment when subjects know they are taking part: biases on the treatment group (Hawthorne effect) or on the control group (John Henry effect). In the medical field, single or double blind (subjects and experimenters) RCTs are the usual way to control these biases (see Abramowicz and Szafarz, Chapter 10, this volume).

demonstrate that it is hard to characterize the microcredit target population. This rules out the possibility of extrapolation to a wider population as a legitimate action. A fortiori, the samples surveyed are not representative of anything aside from themselves: the households surveyed in the case of Bosnia and Mongolia, and the expansion areas (selected villages and neighborhoods) in the case of the other four. Moreover, this property only exists in theory: the multiple failings of the survey protocols in the field mean that the expansion zones' theoretically representative samples are not representative in practice.

If data cannot be extrapolated, comparison with other sources can be instructive to qualify respondent profiles. Official figures from representative surveys conducted by national statistical offices are a good benchmark to characterize a national or local context. Only two studies did so (Bosnia and Mongolia). In the other four studies, it is hard to get any idea of who was surveyed. As reported above, we performed this exercise for the Moroccan RCT. We have shown, among other results, that the average household size is atypical and tends to increase, while it decreases across the rest of the population over the same period. To take this assessment further, we use the typology of hazards to external validity established by Peters, Langbein, and Roberts (2018): Hawthorne and John Henry effects and general equilibrium effects (the others being addressed above). The papers do not discuss these hazards and many do not provide the necessary information to assess potential problems, except (partially) for Hawthorne effects (Bosnia and indirectly Mexico, see the discussion on ethics below) and general equilibrium and spillover effects (Morocco), despite the fact that they are at work in all cases.

Are these external and internal validity threats acknowledged by the authors? More broadly, what types of caveats do the papers mention? We report on them in Table 7.4. With the exception of the Moroccan RCT, the authors discuss a number of caveats. Almost all mention the lack of external validity given the lack of statistical power due to insufficient sample sizes. Heterogeneity to treatment is also widely acknowledged. The fact that the other RCTs return similar (but equally underpowered) results is considered as a source of robustness (see, for instance, Banerjee et al. 2015b: 25). In addition, more specific caveats are quoted such as non-compliance with randomization design (Ethiopia) and selective attrition (India and Ethiopia), and measurement errors (Ethiopia). These observations tend to confirm the persistence of the Peters, Langbein, and Roberts (2018) results concerning the limited attention paid to external validity, to which we should add the internal validity problems raised above.

Lastly, ethical considerations warrant discussion, as this issue is of specific concern to RCTs in general (see the Introduction, Chapter 1 (Ravallion) and Chapter 10 (Abramowicz and Szafarz) in this volume). These considerations are absent from all articles when they should be stressed. The papers do not specify whether the informed consent of the participants was requested and obtained,

with the exception of Angelucci et al. (2015). In addition, the information that they report to have imparted to the participants is partial: they specify, possibly to rule out suspicion of a Hawthorne effect, that they asked for an agreement to participate in a "*comprehensive socioeconomic research survey.*" Yet they knowingly failed to mention that the survey was connected with Compartamos and especially that it was part of an experiment. A look at the available survey questionnaires (Bosnia and Morocco) shows that, in these two cases, respondents were not informed that they were participating in an experiment. The Bosnian RCT raises further ethical issues. This RCT consisted of granting credit to individuals initially rejected by the MFI's risk criteria, as done in South Africa and the Philippines (Karlan and Zinman 2009, 2011). This strategy placed the treated group at risk, at odds with the "do no harm" principle. The RCT confirms that marginal customers have significantly more repayment difficulties than regular customers, with a risk of over-indebtedness.[14]

Now that we have discussed the issues of internal and external validity, we turn to the question of the impacts themselves. Even without considering the limitations outlined above, and sticking to the results proposed by the authors, the impacts are problematic. Table 7.5 provides an overview. First, take-up data are unreliable and often contradictory between survey and administrative sources. The Moroccan case shows that the inconsistencies go beyond differences in averages and under-reporting (see the section The Emblematic Case of the Moroccan RCT on the discrepancies between administrative and survey data). On average, the experiments' impacts on credit take-up range from 8 percent to 50 percent when clusters were randomized to 98.5 percent in Bosnia where individuals were randomized.

Regarding the impacts of microcredit, low take-up has huge implications in terms of the significance of the coefficient. Dahal and Fiala (2020) replicate the six AEJ:AE RCTs. They find that each one is significantly underpowered due to the low take-up of the financial product offered. Even after pooling the data, the minimum detectible effect magnitudes are still very large: 230 percent for main outcomes under perfect compliance and 1,000 percent under actual compliance. They conclude in their abstract that, "*The existing research on the impact of microfinance is generally underpowered to identify impacts reliably and suggests that we still know very little about the impact of microfinance.*" Although Banerjee et al. (2015b) acknowledge the problem of underpower in their introduction, Dahal and Fiala (2020) is the first paper to quantify how big of an issue it is. It confirms the previous study by McKenzie (2012), which estimates the necessary sample size at 15,000,000 to be able to secure the power to identify impact magnitudes of 10 percent in the Indian RCT.

[14] "All this suggests that the loan officers had good reason to classify our target population as marginal" (Augsburg et al. 2015: 201).

Table 7.5 Impact, references, and publications

	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco
Impacts						
MFI credit take-up						
Data source	Survey	Survey	Survey	Admin, survey	Survey	Admin, survey
Presence of other MFIs	Yes	Yes	Yes	Yes	Yes	Yes
Substitution/Crowding effect	n.a.	No	Yes (substitution)	Yes (crowding-in)	Yes (substitution)	Yes (substitution)
Impact	Positive (98.5%)	Positive (25%)	Positive (13%)	Positive 8% (survey), 11% (admin)	Positive (50%)	Positive Survey (9%), 17% (admin)
Outcomes (others than credit take-up)						
#	47	37	99	37	41	37
# of sign. Impact (at 1%; 10%)	0/47 (1%) 3/47 (10%)	0/37 (1%) 5/37 (10%)	1/99 (1%) 13/99 (10%)	3/37 (1%) 9/37 (10%)	0/41 (1%) 10/41 (10%)	6/37 (1%) 17/37 (10%)
References, publications						
# of references:	22	24	27	28	37	16
Of which RCT	5	11	11	20	10	8
Of which methodology/ theory	4	4	3	4	18	6
Of which other microcredit methods	6	7	4	4	5	0
Others	7	2	9	0	4	2
# of papers in academic journals	1 (AEJ:AE)	2 (AEJ:AE; Demography)	1 (AEJ:AE)	1 (AEJ:AE)	1 (AEJ:AE)	1 (AEJ:AE)

Source: Authors based on Banerjee et al. (2015c).

Note: The high number of outcomes in the Indian RCT (99) is due to the fact that two endline surveys were performed. Significant impact (at 1 percent; 10 percent): number of impact estimates associated with p-values significant at the 1 percent level and at the 10 percent level.

Turning to the impacts on the selected outcomes, the presentation on this issue made by the authors of the General Introduction (see Table 7.2), which is supposed to summarize the consolidated results of the six RCTs, is misleading. An exhaustive count of estimated impacts on all variables considered in the six papers draws the following conclusions. No less than 298 impacts are estimated throughout the volume (excluding quantile estimates). Of this total, only 10 are significant at the 1 percent level, meaning that 97 percent of the possible retained effects are not significantly different from zero. Three RCTs have no significant impact at all (Bosnia: 0/47, Ethiopia: 0/37, and Mongolia: 0/41) and one has only one significant impact (India: 1/99). Even when the threshold is relaxed to 10 percent (a more lenient threshold than in usual practice), 81 percent of the effects are not significant. The Bosnian RCT is an extreme case in this respect with only three significant impacts at this threshold out of the 47 tested. Such proportions raise all the more doubt as all the articles mention a systematic problem of statistical underpower, which would explain the lack of impact. The sample sizes are not large enough to estimate the impacts given the low take-up, and this is indeed what we find. Moreover, 60 percent of the significant impacts (at 1 percent) come from the Moroccan RCT, whereas it represents just 12 percent of the total number of estimated impacts. This result confirms the central role played by this experiment in the Special Issue, above and beyond the praise it has attracted for its sampling strategy and spillover estimates. However, we have shown the doubtful nature of the results obtained by this RCT. This further reduces the number of significant impacts, which were already impressively low.

Symptomatically, the transition from academic papers' results to the General Introduction, and then to the synthesis in the Policy Bulletin (J-PAL and IPA 2015) proceeds, by successive approximations, to simplify and magnify the lessons, even to the point of displaying erroneous results. If we go back to the summary of the impacts presented in the Policy Bulletin (Table 2, p. 11; see also our Table 7.2), out of the 48 impacts measured (8 outcomes and 6 countries), 16 are announced as significant (14 positive and 2 negative). That is essentially wide of the mark. First, the significance threshold chosen is 10 percent, which is a level of precision at the upper limit of that which is usually used. If we adopt a more demanding threshold closer to standard practices (i.e. 1 percent), none of the 16 impacts is significant.

A more detailed analysis of the 16 selected impacts finds many inconsistencies. For Bosnia and Herzegovina, the impacts on *Business ownership* and *Business inventory/Assets* are announced as positive. But the first is not significant at 10 percent. As for the second, what is significant at 10 percent is a dummy variable measuring whether the firm owns capital or not. The impact on the total value of *Assets* is negative (although not significant), so at best null. For Ethiopia, the only impact considered significant and negative is that on *Household spending/consumption*. Yet consumption was not measured in the survey. In India, the two

positive impacts are on *Business inventory/Assets* and *Business inventory/costs*. Neither impact is robust: the first impact is positive in the second endline survey, but not significant in the first survey, and vice-versa for the impact on *Business inventory/costs*. In Mexico, two positive impacts are noted. While the impact holds for *Business Revenue*, no data allows for a measurement of *Investment* (the second outcome assumed to increase with the treatment). *Assets* are furthermore decreasing (effect significant at 5 percent). In Mongolia, three outcomes are expected to have positive effects. This conclusion holds for two of them: *Business ownership* and *Household consumption* (at 10 percent). However, although the composite index of *Assets* is positively impacted (at 10 percent), the effect is non-significant (and even negative) for the *Assets* value. In the case of Morocco, where four outcomes are considered positive, we would refer to this RCT's abovementioned reliability issues. The synthesis of the Policy Bulletin appears biased, or at best highly imprecise.

Given these shortcomings, the high but non-significant coefficients would have been the same even if the sample sizes had been sufficient. These results have two implications. First, they place a question mark over the general statement that microcredit is not "transformative." This may be so, but nobody has produced any reliable evidence on this question. Second, Dahal and Fiala (2020) conclude that, "*Existing research… suggests that we still know very little about the impact of microfinance,*"[15]. This paradox in view of the amount of resources put into RCTs on microcredit is confirmed by Jonathan Morduch (2020b), one of the best specialist of microcredit worldwide (*Why RCTs failed to answer the biggest questions about microcredit impact*).

Another finding for both external and internal validity is the fact that none of the replication studies (Dahal and Fiala 2020; Kingi et al. 2018; Meager 2019) pointed up the errors we documented in our Moroccan replication. This includes the most obvious such as the authors' statements about the total absence of contamination in the control groups, the inconsistent household counts before and after trimming, and the claim that no trimming was conducted at baseline. This underlines the shortcomings of "push-button replications" or replications that apply different econometric specifications to the same data without checking the reliability of the original data, codes or sampling.

7.4 Results: From Statistical Biases to Interpretative Biases

Section 3 explores the fabric of RCTs in the field and highlights the many weaknesses of RCTs on microcredit in terms of their internal and external validity

[15] This point is acknowledged in a roundabout way by the editors of the Special Issue: "The individual studies may lack strong evidence for transformative effects on the average borrower, but they also lack strong evidence against transformative effects" (Banerjee, Karlan, and Zinman 2015: 3).

issues. The stages of statistical data collection and econometric analysis are then followed by the stage of interpretation: "*The beauty of randomized evaluations is that the results are what they are: we compare the outcome in the treatment with the outcome in the control group, see whether they are different, and if so by how much,*" (Banerjee 2007: 115–16). An analysis of how randomistas transform their data into scientific statements would appear to challenge this so-called "beauty" of RCTs.

Taken in isolation, most of the six RCTs' econometric results are meaningless in themselves, let alone in the absence of contextual information. The authors, particularly in the General Introduction, make this interpretation in a highly specific context and at the cost of implicit, but strong assumptions borrowed from a behavioral theory of change. Anthropology and political economy frameworks would return very different conclusions. Our purpose is not to disqualify the process of interpretation, which is inherent in data analysis, but to point out that randomistas, contrary to what they claim, cannot escape it. Results are not "what they are," as Naila Kabeer also shows using qualitative tools to revisit a field studied by an RCT (Kabeer 2019).

Moreover, their interpretation is based on a "persuasive rhetoric" (Labrousse, Chapter 8), which consists of making a clean sweep of previous research and extrapolating (here, we find the problem of external validity), while overriding specific issues that are essential to understand the impacts of microcredit, and which other methods have already addressed.

7.4.1 Making a Clean Sweep of Previous Research

Randomistas' results are often presented as unprecedented "discoveries," whereas they are often only the replication of conclusions obtained from previous studies, primarily those obtained from non-experimental methods that are almost never cited (Labrousse 2010). The General Introduction is a good illustration of this. The results are presented as the first scientific evidence of the impacts of microcredit. "The evidentiary base for anointing microcredit was quite thin" (Banerjee, Karlan, and Zinman 2015: 1). Up to this point, available empirical evidence had been based on "anecdotes, descriptive statistics or impact studies that are unable to distinguish causality from correlation" (pp. 1–2). The authors claim to be part of "the debates that took place in the 2000s and continue today" (p. 2) but these debates are actually taking place in a surprisingly cloistered world. Of the 18 references in the General Introduction, 12 (two-thirds) come from the authors themselves and 17 (94.4 percent) from J-PAL members. Only one article escapes this endogamic principle.

No non-randomized studies are cited. Looking at the six articles in the Special Issue, the article on Morocco is equally exclusive (only RCTs are mentioned). The

others are less so, although variably as shown in Table 7.5. The Bosnia and Herzegovina study is the most pluralistic, with an RCT/non-RCT ratio of 0.8; this ratio ranges from 1.57 to 5 for the others.

In addition to the disregard for available non-RCT evidence, there is a tendency to extrapolate and pass over key issues. Without claiming to be exhaustive, but focusing on the points that we feel are key, we address in turn the issues of take-up, business creation and freedom of choice, social transfers and self-reliance, and the problem of over-indebtedness.

7.4.2 Take-up

Low take-up is certainly the most accomplished result of the Special Issue. Many practitioners, decision-makers and researchers, even today, still predict an unlimited market, confusing financial exclusion with demand for credit. Although this result is useful, its true significance is limited. First of all, it should be noted that this exercise is nothing new. Some studies have long warned of low demand (Johnson and Rogaly 1997; Servet 2006), including providing quantitative estimates (Hes and Poledňáková 2013; Khandker, Hussain, and Khan 1998). Moreover, the take-up rates referred to here are difficult to compare and interpret given the diversity of protocols and randomization methods (see "External Validity" section). It is therefore difficult to assess the nature and significance of the target population, and consequently to draw operational conclusions. Moreover, the RCTs say nothing about the reasons behind the low take-up: does it reflect an intrinsically low demand and low propensity to get into debt, and/or does it reflect the inadequacy of the supply, with the two explanations not being mutually exclusive? Only more detailed data could answer this question based on a detailed analysis of financial practices, as seen with financial diaries (Collins et al. 2009) and their social, moral, and political meanings (see, for example, the qualitative analysis of the Moroccan context, overlooked by the authors of the Moroccan RCT; Morvant-Roux et al. 2014).

7.4.3 Microcredit, Self-employment and Freedom of Choice

The six studies in the Special Issue tend to concur that there are limited impacts on the creation of new businesses (significant only in two cases), with the expansion of existing businesses being more frequent (four cases). Improved profitability is found in only one case (Morocco), but we have seen above the low internal validity of these results. Moreover, even when a business is started up or expanded, no impact on income growth is observed, either because profitability is low or because self-employment income is offset by a decrease in paid work elsewhere.

The authors of the General Introduction thus claim to draw a novel conclusion on the impact of microcredit on entrepreneurship.

However, since the late 1980s, numerous empirical studies have been conducted to measure the impact of microcredit.[16] The systematic review by Duvendack et al. (2011), conducted when RCTs were just starting to be used, draws two conclusions. First, a large number of quantitative studies, both experimental (including RCTs) and observational, are subject to multiple biases.[17] Second, when results are valid, they reveal a limited and heterogeneous impact, something that Morduch also observed in the late 1990s in his pioneering article on the partly unfulfilled promises of microcredit (Morduch 1999). So the results of the Special Issue do not look so new after all. More importantly, given the complexity of the causal chains induced by microcredit (Duvendack et al. 2011) and the heterogeneity of effects and types of microcredit,[18] RCTs do not seem appropriate (Bernard, Delarue, and Naudet 2012). Ultimately, the randomistas' question—Does microcredit work or not?—is poorly placed. What is shown by rigorous studies (whether quantitative, qualitative or mixed) is that certain types of microcredit may be useful for certain categories of populations and in certain contexts, but not others (Bédécarrats 2012; Copestake et al. 2016). For instance, the work of Copestake et al. in Peru and Zambia (Copestake, Bhalotra, and Johnson 2001; Copestake et al. 2005) and Bouquet et al. in Madagascar (Bouquet et al. 2007) shows in detail which categories of populations benefit from microcredit and why and, conversely, which categories see their situation deteriorate, with direct operational conclusions on how to transform the services offered. Still in Madagascar and ten years before the Special Issue, our own impact evaluation of a local MFI based on a quasi-experimental approach suggests three main stylized features presented later as "discoveries" by the randomistas: the impact of microcredit is not "transformative"; the impacts are heterogeneous across the firm size distribution; and context matters: microcredit is more beneficial in time of growth than in time of crisis (Gubert and Roubaud 2011).

[16] Bédécarrats (2012) identified 154 impact studies, compared with 51 for Duvendack et al. (2011).

[17] Several replications of non-experimental studies long used as "evidence" of the positive impact of microcredit have revealed numerous biases and an overestimation of impacts. See Duvendack and Palmer-Jones (2012); Roodman and Morduch (2014).

[18] We shall give the example of rural microcredit, which is widely represented in the Special Issue. Over and above the credit modalities, what are the credit needs (inputs, equipment, livestock, cash flow to finance the lean season, etc.); what type of agriculture are we talking about (cash or food crops, agriculture in dry or rainfed areas, intensive or extensive, family-based or professional, independent or contractual through integration into agro-business sectors or producers' cooperatives, etc.); and what is the nature of the rural economies (degree of monetarization, remoteness and quality of infrastructure, non-farm income opportunities)? Last but not least, what kind of MFIs are we talking about? Status (for-profit/not-for-profit) is one thing (specified in the Special Issue), but other key questions include mode of governance, degree of integration and adaptation to local realities, and capacity to design products adapted to local demand. In view of this diversity, it makes no sense to talk about "rural microcredit." On this diversity, see for example (Morvant-Roux 2009).

Understanding the heterogeneity of impacts (and drawing operational conclusions) requires a different conception of causality mechanisms, not in terms of "difference-making" but in terms of "mechanism" and "process" (Shaffer 2015). Moreover, given the many externalities, focusing on individual impact is also restrictive. Very few studies have applied general equilibrium models to the case of microcredit at mesoscale (for one exception, see Mahjabeen 2008). Examinations of externalities have been carried out mainly by political economy analyses, considering that it is precisely the analysis of the embeddedness of MFIs in their social, cultural, political, and economic environment and externalities that has a powerful effect on product uptake and hence impact (Copestake et al. 2016). Convincing and useful impact studies in rural areas have shown the key role of financial innovations anchored in local territories—capable of developing specific products designed locally (bridge loans, guarantee funds and leasing) and combining with other measures (cropping contracts, harvest warehouse, technical assistance, etc.)—in enabling small farmers to upgrade their participation in various value chains (Bastiaensen and Marchetti 2011; Bouquet et al. 2007), while often encountering threshold effects (Doligez 2002). Effects are sometimes questionable, such as when microcredit accelerates migration processes, as migration is necessary to repay microcredits (Bylander 2014; Morvant-Roux 2013). They are sometimes more political and cultural than economic in nature. For example in Egypt, the introduction of microcredit disrupts local values—understood broadly as what makes sense to people—and thereby the processes of recognition, identity and socialization (Elyachar 2006). In rural South India, the massive presence of MFIs in certain territories reconfigures local power relations and chains of patronage by feminizing them (Guérin and Kumar 2017). These results (and their related questions) are far removed from those of the randomistas. And yet, if we really want to understand what microcredit is changing in people's lives, it is precisely these kinds of broad questions that need to be asked.

In addition to these in-depth studies, which are systematically based on sound knowledge of local contexts over time, it is useful to mention other, lighter methods designed to quickly identify the characteristics of customers (and non-customers) and the way in which services are used, and to derive recommendations for improving the quality of supply, which remains the key recurrent question asked by microcredit providers.[19]

We shall now come back to the Special Issue. Not only do the authors not bring anything fundamentally new to the existing evidence, but their interpretation of the quantitative results is problematic. Microenterprise may reflect absence

[19] Examples include the tools developed by AIMS (Assessing the Impact of Microenterprise Services) and Imp-act, which have been denigrated for their lack of a sophisticated quantitative method. These tools may have lost their relevance to "prove" impact on a large scale, but they have nevertheless been very useful to "improve" and diversify the microfinance service supply.

rather than expansion of choices. A large proportion of micro-entrepreneurs, condemned to self-employment for lack of paid employment, resemble more the self-exploitation analysed by Alexander Chayanov (1966 [1925]) than the Schumpeterian entrepreneur (Lautier 2004). The case of Mongolia is instructive in this regard. The RCT shows that access to group credit allows women to start new micro-enterprises, but for negative incomes, while their working time increases by more than a third (without any change in household time). These negative effects are mainly observed for less-educated women (Attanasio et al. 2015: 105, note 21). The authors believe that profitability may improve once the credit is repaid (Attanasio et al. 2015: 115). Here, we find the problem of temporality, already highlighted as a strong limitation of the RCT (Bédécarrats, Guérin, and Roubaud 2019; Labrousse 2010). These women may indeed have chosen to embark upon the entrepreneurial venture, and this may explain the improvement in consumption (results indicate more and healthier consumption). But what is the meaning of this 'choice' and, above all, what are its consequences if it then gives rise to increased responsibilities and possibly disengagement by other household members (and hence intra-family inequalities)? The quantitative data do not enable a conclusion to be drawn, and the authors of the RCT do not make any particular judgement. A robust interpretation would call for other types of data, quantitative or qualitative. The authors of the General Introduction, on the other hand, focus only on the "freedom of choice" dimension, without mentioning the potentially negative effects of these "choices" on women, especially the most disadvantaged.

7.4.4 Microcredit, Social Expenses, Social Transfers and Self-reliance

While the effects in terms of business and income are inconclusive, the authors of the General Introduction observe what they describe as positive effects on two indicators: "non-essential expenditures," a sign of better discipline and management skills, and a decrease in "social transfers," a sign of greater autonomy. "Non-essential expenditures" include "temptation goods" and decreased in four countries (they were not measured in Ethiopia and the results were not significant in Mongolia): alcohol and cigarettes in Bosnia-Herzegovina; cigarettes, sweets and soda in Mexico; and alcohol, tobacco, betel leaves, gambling and food consumed outside the home in India. These expenditures also included festivals, with decreases observed in India and Morocco.

The authors put forward a number of explanations to explain this decrease in "temptation goods": repayment and investment constraints, better self-discipline, and more involvement of women in decision-making. The reduction in temptation goods is one of the major results of the Indian RCT, highlighted in the abstract. The study's authors take care to specify that it is the populations

themselves who describe these goods as "temptation goods," in the sense that they would like to reduce them (Banerjee et al. 2015b: 24). But for people to express this preference (an observation of vague origin, seeming to be more a matter of "anecdotes" whose use is highlighted by A. Labrousse, Chapter 8) may well reflect that they have taken on board the moralizing discourses frequently given by development organizations (including MFIs), and this since the colonial period.[20]

Moving beyond the moralizing dimension of the randomistas' conclusions,[21] a detailed analysis of the meanings and role of these outlays could shed a different light. On the subject of alcohol, no one would dispute that excessive consumption is a public health concern. Yet if we really want to understand this type of consumption and devise courses of action, it is essential to recognize the social and political dimension of alcohol. Like many other temptation goods, and contrary to what behavioral economics suggests, it is not a good defined solely by its "immediate utility" (Banerjee and Mullainathan 2010). Alcohol can play a social role since it enables workers to endure physical work and access socialization spaces and therefore strategic information (bars are often strategic places to negotiate employment contracts and orders; Picherit 2018). Alcohol can play a political role when it gives workers the opportunity to make demands of employers and bosses that are more easily acceptable under the influence of drunkenness (Scott 1977). Above all, alcohol is frequently deliberately offered by employers and labor recruiters in order to build loyalty (Picherit 2018). To think that sacrifice or more self-control would be enough to fight these "temptations" is therefore fallacious.

Similarly, catering expenses (meals and tea) outside the home are not solely "lucrative opportunities to save" (Banerjee and Duflo 2011: 170). Street restaurants and tea shops are eminently strategic places. In an informal opaque economy, structured by interpersonal relationships, these spaces enable traders to keep each other informed of the market situation, price trends, opportunities to be seized, possible sources of financing, risks of tax or police checks, etc. Small entrepreneurs cultivate exchange and mutual support links, whose role is often decisive for the survival of their business.

On the subject of expenditure on social and religious rituals, anthropology has long shown that "social wealth" is an essential factor of success and protection (Guyer 1997) and that "investing" in social relations can, in certain situations, be

[20] In India, for example, reports from British settlers and Christian missions in the early nineteenth century already mentioned the improvidence and prodigality of the poor (Cederlöf 1997; Hardiman 2000).

[21] The statements made by the randomistas are reminiscent of the Victorian morality of the European industrial revolution, legitimized by the arguments of neoclassical economists of the time. Faced with the extreme poverty of the working-class world during the British Industrial Revolution, some lamented the poor's lack of self-reliance, lack of foresight and wasteful alcohol expenditure, and argued for financial education courses rather than wage increases (see for instance (Jevons 1883: 196–200; 205).

much more rational than trying to save money by cutting oneself off from one's surroundings (Narotzky and Besnier 2014). Beyond randomistas, the question of "community taxes" and their cost benefits in terms of protection has been the subject of various studies by development economists. But these studies rarely take into account the complexity of the financial channels to which these expenses give rise and their long-term nature. An analysis conducted in India of the correlation between festival expenditure and lunch invitations shows, for example, that these expenses act as safety nets (Rao 2001). Moreover, what is considered by economists as an expense is sometimes conceived as an entitlement or as savings, since it will give rise to a future counter-gift. Also in India, accounting for all debts and entitlements generated over time by ceremonial spending, which families are well aware of because they calculate in these terms, shows that families' net financial wealth is radically different from that suggested by an analysis in terms of "spending" (Guérin, Venkatasubramanian and Kumar 2019). This contradicts the short-term bias that randomistas often attribute to the poor (Banerjee and Duflo 2011: 183–204).

With regards to social transfers, of the eight estimates retained (which concern transfers from the family or the State), five are negative. This observation leads the editors of the Special Issue to conclude that "self-reliance" has improved, a factor that is judged in a positive light.[22] This interpretation is both risky—there is no reason to believe that the decrease in transfers from family and friends is seen as positive or a source of well-being by the people concerned themselves—and normative, as are previous interpretations. Here again, anthropology is valuable in elucidating the decisive role of social interdependencies, in terms of both material protection and identity. Looking past the randomistas, there are those in the development world—policymakers, practitioners and some researchers—who consider dependency both as a political problem (assistance is expensive) and as a moral problem (dependency is seen as being incompatible with individual freedom). However, in many contexts, being connected and dependent on others is both a mode of action and a deliberate strategy. Rather, people's agency translates into the ability to choose certain forms of dependency and interdependence.[23]

Ultimately, the General Introduction's conclusion on improving self-reliance, as well as that on "freedom of choice," is driven by specific interpretations of econometric results (if not extrapolations from the conclusions of some of the RCTs). These interpretations are underpinned by a singular conception of

[22] It should be noted, however, that this interpretation is that of the authors of the introduction, and not of the authors of the papers, who either do not comment on this result or underline its ambiguity. On Mongolia, the authors mention, for example "Increased within-group financial discipline may come at the cost of disrupting informal credit and insurance systems based on kinship and other social ties" (Attanasio et al. 2015: 114).

[23] For a general overview of how anthropology addresses this issue, see for example Ferguson (2015).

individual autonomy and freedom, and thereby their own theory of change, viewing people as isolated atoms, denying the multiple roles that social interdependencies play at different levels and implicitly considering these interdependencies as harmful. These two conclusions—"self-reliance" and "freedom of choice"—were nonetheless included in J-PAL and IPA's Polict Bulletin (J-PAL and IPA 2015), which was then widely disseminated by many blogs and discussion networks and seen as an indisputable asset of this research.

7.4.5 Microcredit and Over-indebtedness

A major conclusion of the Special Issue is that microcredit is not the "debt trap" denounced by microcredit opponents. First of all, it should be noted that no scientific studies are mentioned in the General Introduction, as if the "debt trap" were anecdotal evidence. It is true that when a number of microcredit repayment crises erupted, the media made a big deal of it (in the same way as they had praised microcredit when it started). However, the press aside, there is a vast body of scientific literature dealing with household over-indebtedness in the Global South and the role played by microcredit,[24] including in the countries covered by the Special Issue. A number of problems arise here.

The first concerns external validity, where extrapolation occurs without taking into account the singularity of the contexts studied and the fact that this is microcredit "on the margin" (Wydick 2016). The six RCTs focused on areas and populations that were supposed to be free of microcredit.[25] However, by definition, the problem of over-indebtedness is less acute than in areas and populations previously exposed to microcredit. It is therefore tautological that the "debt trap" does not appear. Yet over-indebtedness among some of the microcredit clients has been documented and sometimes measured in four of the countries studied.[26] The fact that the RCTs did not quantify it does not enable them to conclude that the debt trap does not exist. Contrary to what the authors of the General Introduction suggest, the available literature is not content to make do with "anecdotes." Scholars demonstrate (most often qualitatively) the role of microcredit based on a detailed analysis of its specific characteristics in relation to other sources of debt, in particular the rigidity of the repayment terms and low tolerance for non-payment. In some contexts and MFIs, this zero tolerance

[24] In addition to the references already mentioned, see (Schicks 2013; Schicks and Rosenberg 2011; Guérin, Morvant-Roux, and Villarreal 2013; Guérin, Labie, and Servet 2015).

[25] As mentioned above, this virginity was in fact a decoy and all control populations actually had access to microcredit. However, the market was not saturated as it might have been elsewhere, so there was less of a risk of over-indebtedness.

[26] For Mexico, see Morvant-Roux (2013), Angulo Salazar (2013), Hummel (2013), Rozas (2014). For India, see Guérin et al. (2013), Joseph (2013), Taylor (2011), Prathap and Khaitan (2016). For Bosnia-Herzegovina, see Maurer and Pytkowska (2011); Opem and Goronja 2013; Bateman 2010). For Mongolia, see Javoy and Rozas (2013).

takes the form of coercive enforcement procedures.[27] These scholars also propose a nuanced and contextualized analysis, highlighting the role of the global context (including stagnant and declining real incomes in the face of growing needs) as well as the ambivalent role of microcredit (for some borrowers, microcredit can be a way to repay informal debts and reduce over-indebtedness).[28] The causal link between microcredit and over-indebtedness may only concern a minority of microcredit clients (which brings us back to the issue of heterogeneity). But its repercussions (impoverishment, social exclusion, suicide, etc.) (Schicks 2013) are sufficiently tragic to warrant randomistas taking the phenomenon more seriously.

The second problem is the extrapolation from the six case studies by the introduction's authors. Even for areas and populations recently exposed to microcredit, over-indebtedness cannot be ruled out. The Bosnia and Herzegovina RCT was conducted in a context of a proven over-indebtedness crisis, which the authors mention as contextual data (Augsburg et al. 2015: 185). This RCT specifically concludes that the treatment group had repayment difficulties (Augsburg et al. 2015: 199–201), and that these repayment difficulties are a potential symptom of over-indebtedness.[29] The RCT does not enable a conclusion of either the existence of over-indebtedness or the role of microcredit. However, the existence of a "debt trap" cannot be excluded. In the Mongolian RCT, the authors take care to specify that their study does not measure over-indebtedness, but only repayment defaults, which are two distinct things.[30] The special issue's introduction makes no reference to these clarifications.

In Morocco, a qualitative study conducted by one of us at the same time as the RCT concluded that there was low propensity for debt in rural areas, for cultural and religious reasons (Morvant-Roux et al. 2014). This general observation, valid "on average," does not, however, exclude over-indebtedness problems among a fraction of the population. Given that Morocco also experienced a default crisis (which the authors do not mention, although it took place during the RCT), MFIs concentrate their supply on a minority of clients judged solvent and reliable. These clients are hence overexposed to microcredit, and some of whom do face over-indebtedness problems (Morvant-Roux and Roesch 2015).

[27] In India, for example, the prosecution of defaulters in the workplace or at home, public denunciations and insults, solicitation of relatives, physical threats, confiscation of property and administrative documents; in some cases, the most recalcitrant have been tied up in a public square or in direct sunlight (Arunachalam 2011; Servet 2011).

[28] As we finalize this chapter (October 2019), the United Nations has just taken up this issue, commissioning a report on the subject. This would seem to suggest that the problem does exist. https://www.ohchr.org/EN/Issues/Development/IEDebt/Pages/ReportPrivateDebt.aspx

[29] Defaults can also be "strategic" defaults expressing a refusal to repay, particularly in the context of a repayment crisis.

[30] Since some defaults can be strategic, good repayment rates can mask sacrifices made to honor debts, which the authors of the RCT in Mongolia acknowledge (Attanasio et al. 2015, footnote 25, p. 114).

Like Bosnia and Herzegovina, India has been hit by some major microcredit default crises: in Krishna District in Andhra Pradesh back in 2006, then in a small town in Karnataka in 2009, and in the entire state of Andhra Pradesh in 2010. Analyses of this crisis, both quantitative and qualitative, have highlighted the existence of an over-indebtedness problem for some of the clients. The over-indebtedness of poor populations, with or without microcredit, has also been documented outside of default crisis areas, including in urban areas. As already mentioned, the Indian RCT was conducted from 2005 to 2010 in marginal areas of Hyderabad newly exposed to microcredit. Yet how is it possible to extrapolate from this highly specific case study when there is a vast body of evidence demonstrating the existence of over-indebtedness? On this issue, the article by Banerjee et al. cites just one press article, "Anecdotes about highly successful entrepreneurs or deeply indebted borrowers tell us nothing about the effect of microfinance on the average borrower, much less the effect of having access to it on the average household," (Banerjee et al. 2015b: 23). In view of the state of the art's alert over the level of over-indebtedness among poor Indian populations, and given the extreme specificity of the districts they study, is it not maybe their own study that should be qualified as anecdotal?

Finally, the question may be put as to whether the measurement of household debt was properly conducted. The collection of reliable debt data calls for a number of precautionary measures for the following reasons: debt taboo, exacerbated when MFIs claim to eradicate informal borrowing since this encourages clients to conceal their informal debts; diverse terminology used; and the range of debts that may be held by different family members without their necessarily sharing that information. Given the approximations observed at the other stages of data collection and analysis (see Section 7.3), it is not unreasonable to question the ability of the randomistas to design a questionnaire that can adequately capture household debt. However, it should be noted that this difficulty is not unique to the randomistas. Collecting reliable data on incomes in the Global South has taken decades of learning to adapt the statistical tools to contexts where households juggle different sources of income, including informal sources. The same work has yet to be done on debt, which remains poorly measured and often underestimated.

7.5 Conclusion and Discussion

Given the many limitations and shortcomings we have found with the method, applied here to microcredit, the question could be asked as to why RCTs have had such academic, media, and political success. We have already explored the reasons for this contradiction (Bédécarrats, Guérin and Roubaud 2019) in a study of the political economy of what has now become a real industry (see Ravallion,

Chapter 1 in this volume). As with any industry, the impact evaluation market is where supply meets demand. We have explored these two elements in detail, showing that the demand is twin-engined, driven by both the donor community and the academic world, while the supply is largely shaped by a brand of scientific businesses and entrepreneurs who appear to have created a new business model designed to build a monopoly and a rent position on the evaluation impact market. Further illustrations of this would-be domination strategy are turned up when exploring how the data have been produced and analysed, as we have done here. In addition to making a clean sweep of the past (see Section 7.4.1), three other strategies appear to be key: disengagement from a "data culture," ignoring criticism (up to a certain point) and sidestepping certain rules of scientific ethics.

7.5.1 Disengagement from a Data Culture

The many data collection and data entry errors observed in the Moroccan RCT would appear to suggest a certain lack of experience and knowledge, as if the purely technical skills required in the second stage (econometrics: addressing bias issues, selection, and identification of a counterfactual) excused the researchers from the need for the know-how required for the first stage (collection of good quality data). To what extent does this concern apply to other RCTs? Unfortunately, that question remains open for the moment, since only full replications would be able to provide the answer. What is clear, however, is that randomistas tend to disregard the debates regarding data collection (as they do the issue of ethics, see Abramowicz and Szafarz, Chapter 10). In most quantitative empirical research protocols, there is a division of labor between data collectors and analysts: the former are statisticians, the latter are economists (econometricians or thematicians). With few exceptions (Deaton 1997; Grosh and Glewwe 2000), few people can occupy both ends of the spectrum. These are fully fledged jobs, requiring distinct skills and training. Statisticians are responsible for the accuracy of the measurement, economists for its relevance, its analysis and the relations and interactions between data. Both activities are essential for the final production of reasonable results, even if statisticians have less social prestige than economists (Desrosières 2013a). Given the skills involved and the way academic journals work, all efforts are concentrated upstream on designing a "smart" randomization process, and downstream on econometric estimations of the impacts with a view to publishing papers in top-ranking reviews.

The disconnect between researchers and the field is another illustration of the data culture. This disconnect is particularly acute at J-PAL. Its hierarchical organization makes for a strict division of labor between project managers, doctoral candidates and field staff (supervisors and investigators). The latter are ultimately given considerable responsibility for which they are arguably not adequately

trained (Jatteau 2018). This division of labor is a practice frequently found in the field of natural and life sciences, but it does not prevent team leaders from staying in regular contact with the data production chain, including for in vivo experiments. Moreover, teams are required to adhere to precise protocols to validate the rigor of the experiments conducted. This cannot be not the case here given the dozens of RCTs in which the most prominent RCT leaders are involved (Bédécarrats, Guérin, and Roubaud 2019). This disconnect has been exacerbated by J-PAL's exceptionally rapid expansion, as mentioned above.

This growth, combined with highly centralized governance, implies that a handful of researchers head up a considerable number of experiments. This in turn places a question mark over their actual capacity to work on each separate RCT (and deepens the disconnect with the field). In February 2019, Esther Duflo had 64 RCTs to her name, equal to just over four new RCTs a year. Dean Karlan, however, is by far the most prolific with 100 trials (and 42 ongoing; January 2017). So how much can they really personally put into each of the RCT results they sign? In fact, the signature of a top randomista researcher appears to be more of a seal to facilitate publication in a top-ranking journal, as part of a global randomista strategy, than a guarantee of research quality.

7.5.2 Ignoring the Critics

Whereas randomistas have built a universal narrative on the impact of microcredit based on this Special Issue (and subsequent publications), other players have drawn different conclusions from these same studies (see also Kabeer 2019). Here again, the Moroccan RCT is a typical illustration. As early as 2009, while the endline was still in progress, the RCT's funder started publicly sharing its feedback on RCTs based on the Moroccan RCT and another study conducted in Cambodia at the same time. The conclusions were clear: they highlighted the challenges faced by the method to produce rigorous impact evaluations given the multiple breaches of protocol that the funder's research team had partially identified (problem of representativeness and product change) and the time constraints that compelled a focus on the short term. Although the findings of the funder's research team have been publicly presented and published on numerous occasions (Bernard, Delarue, and Naudet 2012), they have gone unheeded by the RCT team (Bédécarrats et al. 2019b).

Our own experience with the Moroccan RCT, although illustrative, is a good example of what might be a strategy to ignore the critics, up to a point. In the course of our critical research on RCTs in development, we have invited some of the most vocal RCT proponents to engage in a scientific debate (controversy) on many occasions (dedicated sessions at international conferences). To date, we have received no answer. We also invited ten of the most famous randomistas to

take part in this collective book to balance out the voices on RCTs. They all declined. Directly on the subject of our critical review of the Moroccan RCT, we informed the authors of the completion and publication of our replication (Bédécarrats et al. 2019a). At the same time, we drafted a *Comment* and suggested that AEJ:AE publish the piece with an *Answer to the Comment* from the authors, as is common practice in many journals. AEJ:AE turned down the offer on the basis that the journal does not publish comments. Lastly, when our paper was picked up by coverage in vocal blogs and the press, Crépon et al. (2019) produced a (51-page) Rejoinder using sophisticated analyses to argue that their original results were robust: double post lasso procedure, Benjamini-Hochberg False discovery rate correction of multiple testing, the Bayesian hierarchical model and machine learning analysis, among others, concluding our replication was not scientific. They posted the Rejoinder on their website and enjoined us to post it on the DIAL website, which we duly did. They also informed the AFD hierarchy. IREE suggested both parties publish a short version of the Rejoinder with our answer (Rebuttal of the Rebuttal; Bédécarrats et al. (2019c). In view of the totally contradictory conclusions of the two pieces, we suggested seeking a third party assessment that would decide on whether to retract our replication (Bédécarrats et al. 2019a) or the initial paper (Crépon et al. 2015), depending on the conclusion. Again, they declined the invitation. These episodes illustrate two characteristics of the randomistas' make-up. First, contrary to one of the main selling arguments for RCTs (the simplicity of the method, compared to the so-called "black box" of alternative econometric methods), this type of RCT is extraordinarily complex. In their Rejoinder, they added complexity to the already complex randomization design (which is one of the three paradoxes we sought to explain in Bédécarrats et al. (2019b)). Second, they sidestepped scientific standards by not providing their codes, turning down a peer review of their Rejoinder, and ultimately eluding a fair scientific controversy.

7.5.3 Circumventing Scientific Ethics

In addition to disregarding all things non-RCT, the randomistas have bypassed certain basic rules of scientific conduct. This problem appears to be growing in the scientific community as a whole (Heckman and Moktan 2018). Yet while it is not specific to J-PAL or the randomista community, it is particularly patent here. In the research world, knowledge validation is based on the "peer review" principle, that is a collective review by researchers who critically and anonymously judge the work of their peers. Yet, for this to happen, numerous ethical rules need to be respected, starting with the management of conflicts of interest between authors and members of journals' editorial boards. Editorial favoritism is a recognized and demonstrated process, particularly among economists (Fourcade,

Ollion, and Algan 2015). The Special Issue is illustrative in this regard. The issue's three scientific editors are members of J-PAL (Banerjee, Karlan, and Zinman 2015). In addition to the General Introduction, each editor co-signed an article and two of them were members of the board of editors (Banerjee and Karlan). Esther Duflo was both the journal's editor (and founder) and co-author of two of the six articles. Given, in addition, that nearly half of the articles' authors (11 of the 25) are also members of J-PAL and four others are affiliated professors or PhD students with J-PAL, the journal strayed somewhat from the peer review principles supposed to govern scientific publication. This single example shows in cameo the extraordinary density of the links between RCT promoters identified by Jatteau (2016).

7.5.4 What Remains of the Special Issue?

At the end of the day (or of our in-depth investigation), what have we learned from RCTs on microcredit in the development field? Going back to this chapter's title, if microcredit is not a miracle, as defended by the Special Issue, what are RCTs on microcredit: *miracle or mirage*? Let us wrap up our results and provide key takeaways.

We will start by addressing the internal validity claims, the acclaimed strong points of RCTs. First, as acknowledged by randomistas themselves, there is a lack of strong evidence that microcredit is transformative, just as there is a lack of strong evidence that it is not (Banerjee, Karlan, and Zinman 2015). Given that RCTs are generally underpowered due to low take-up and compliance, we simply do not know. Second, and again acknowledged by the randomistas, heterogeneous effects may be the norm. Microcredit may be transformative for some and not for others (or worse, microcredit may be negatively transformative). Again, given the general underpower of RCTs due to low take-up and compliance, we simply do not know. Furthermore, we do not know why some may benefit from microcredit and some may not (or may suffer a "transformative" penalty). We have no idea through which channels microcredit might have an impact. Third, poor data quality and measurement errors may prompt reconsideration of some of the results that have hitherto been taken for granted. In this respect, the many problems we have identified with the Moroccan RCT need to be taken seriously. Maybe the Moroccan RCT is a one-off (the bad apple). But in this case, its conclusions should be definitively revoked. This would have two direct repercussions. The overall demonstration would be weakened. The "fairly representative sample" used to draw general conclusions would become "less fairly representative." Its good properties put forward in the issue to estimate spillover issues and predict take-up rate, and its sampling strategies to address the issue of low compliance and underpower would evaporate. Maybe it is not a one-off (although we

presume that other RCTs could not perform as poorly), in which case we have a structural problem here. The only way to know would be to conduct full replications, such as ours. We strongly advocate this avenue of research. Fourth, we have shown that many interpretations of the impact of microcredit, underlying the theory of change, are biased, while some obvious impacts (or causes of low take-up) are not even considered. Additionally, other generic concerns remain such as general equilibrium effects, macro policies, etc. (both are internal and external validity concerns).

Second, external validity has never been the RCTs' strong point. Our assessment does nothing to change this view. The usual criticisms, not worth quoting again here, still hold. The Special Issue's novelty is that it considers different RCTs on microcredit taken together in tandem. However, the accumulation of individual cases does not solve the problem. What is gained from diversifying geographic, but hyper-specific contexts, is lost from increasing the heterogeneity of treatment, implementers, and so on. One type of product may work in one context and not in another. Changes to products and allocation schemes do not tie in with "real world" conditions. Lastly, ethical issues remain largely unaddressed despite major departures from good practices in the medical field and even social RCTs in developed countries.

Taking all that into account, what is left? To paraphrase Banerjee and Duflo (2011) as quoted by Agnès Labrousse in her chapter, we can follow up, nearly ten years and dozens of RCTs on microcredit later, with, "Unfortunately, [. . .] ~~until~~ *even* very recently, there ~~was~~ is in fact very little evidence, either way, on these questions. What ~~CGAP~~ *randomistas* call~~s~~ evidence turns out to be case studies [. . .]." Although it is not clear what is left at this stage, what is not left is the huge amount of money and resources spent, some which were withheld from other alternatives and uses. Is it worth spending millions of dollars in return for one single academic paper for each RCT (Table 7-5)? Wouldn't it be more useful for the same sums to be used to fund a developing country's public statistical system to collect a huge amount of representative observational data in the long run? Although RCT proponents have acknowledged some of the methodological shortcomings discussed in this chapter, their answer to resolve them is, "More RCTs!" Yet if RCTs have not delivered on their promises, or at least the promises that randomistas have been selling the world these past two decades, then it would be just as legitimate to say, "No more RCTs!"

We may come across as extreme. Yet the randomista tidal wave has been so powerful (as seen from the way they have swept aside the past by (apparently) ignoring all non-experimental studies) that a small push back in the other direction would do no harm to rebalance the state of the art. Our purpose is not, however, to discredit the RCT method, but to recognize its true value by challenging the pedestal on which it now stands. Rather than "No more RCTs," our advice is actually, "No more standalone RCTs." While RCTs are likely to remain appropriate and

legitimate for certain precisely circumscribed policies, they should still be conducted by the book. Furthermore, they are never self-sufficient. It is both necessary and possible to use other methods without compromising scientific rigor. As we have seen here, this pluralism should be a requirement, in particular to round out RCTs by contextualizing them, both before data collection and for analysis. Pluralism is also a requirement for all development issues, projects and policies not suited to RCTs, and microcredit with its relatively closely targeted interventions is a good example of this given the low take-up and complexity of its effects. Unfortunately, for many RCT proponents, and J-PAL in particular, "RCTs are not just top of the menu of approved methods, nothing else is on the menu," (Ravallion, Chapter 1, this volume).

Acknowledgement

We thank the participants of the March 2019 workshop, which brought together most of the contributors to the book, as well as Solène Morvant-Roux, Jonathan Morduch, and Martin Ravallion for their comments on an earlier version of the chapter.

8

The Rhetorical Superiority of Poor Economics

Agnès Labrousse

8.1 Poor Economics: A Puzzling Success, a Persuasive Rhetoric

As McCloskey (1983) has shown, rhetoric is indeed omnipresent in economics. However, this discursive dimension is often masked: there is a credo of scientific objectivity associated with the rigor of numbers. The *Poor Economics* of the J-PAL are no exception to this rule: their claim to legitimacy relies on the hard numbers produced by randomized control trials (RCT) and the use of anecdotal evidence is explicitly prohibited. Compared to other mainstream economic trends, this discourse adopts particular rhetorical characteristics. As we will see, it is precisely these characteristics that make it persuasive and contribute to the success of this lab and of randomization among diverse audiences. This tremendous success remains puzzling. Despite important and well-identified limits in medicine (Labrousse 2010), in economics (Bédécarrats, Guérin and Roubaud 2019; Pritchett (Chapter 2, this volume), Ravallion (Chapter 1, this volume) and evaluation studies Bernard et al. 2012; Picciotto, Chapter 9, this volume), the boom of RCTs is continuing unabated and the "bubble" has not burst yet. We argue here that the canny rhetoric of J-PAL is a key piece of the puzzle.

We will focus here on the celebrated book *Poor Economics* by Banerjee and Duflo.[1] It condenses the rhetoric of J-PAL—the world-largest lab working on poverty and using RCTs—and its claim to "fight poverty with hard numbers." Hence, it is representative of the dominant usage of RCTs in economics[2] and of its justification. The book is intended for a wide audience, going beyond the academic world to reach members of international and governmental agencies, of NGOs as well as journalists, students, and concerned citizens. It condenses the

[1] Unlike *Poor Economics*, their latest book, *Good Economics for Hard Times*, opens up a broader horizon and, thus, doesn't rely exclusively on RCTs: it refers to many publications and topics that are outside the J-PAL's work.

[2] J-PAL dominates the field (Jatteau 2016). In June 5, 2019, it has conducted 952 RCTs while the total number of registered RCTs in social sciences amounted to 2552, i.e. 37.2% (https://www.socialscienceregistry.org/). Note that other usages of RCTs exist in social and medical sciences (Labrousse 2016).

Agnès Labrousse, *The Rhetorical Superiority of Poor Economics* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0010

results of several hundred experimentations on a multiplicity of topics. The subject of the book is "how to fight global poverty" and "how the poor really live their lives" (p. 14).[3] *Poor Economics* has been subject to much academic and media coverage. Lauded by renowned economists as diverse as Robert Solow, Amartya Sen, William Easterly or Anne Krueger; philanthropists like Bill Gates; quality newspapers (the *New York Times*, *The Guardian*, etc.) as well as business journals: *The Wall Street Journal* praised the book and it received the business book of the year award from the *Financial Times* and Goldman Sachs. It has led to an impressive number of reviews and academic citations (2713 according to *Publish or Perish* as of June 4, 2019). In the US, the book was published by *Public Affairs*. As noted (on p. 297), the owner of this publishing house has also published "Gandhi, Nasser, Toynbee, Truman and about 1,500 other authors."

Reading *Poor Economics,* it seems that the book is flooded with numbers, but also more surprisingly, with anecdotes. In fact, the argumentative efficiency of the book comes in different forms of storytelling, more than through the mere power of naked numbers. The goal of this analysis is (1) to establish, through a textual study, the presence and rationale of salient rhetorical processes, (2) to analyze their effects of persuasion and knowledge, (3) to enlighten both the success of RCTs and some limits of this trendy technique—notably to what issues it makes us blind.

8.1.1 Theoretical Framework: Workaday Rhetoric, Epistemic Communities and Discourse Analysis

To do this, this analysis draws upon the field in economics opened by McCloskey's seminal article (1983). Centered on economists' *workaday rhetoric*, this approach incorporates statistical arguments and formal models in the analysis of rhetorical processes. Unlike McCloskey, rhetoric here is not envisioned as an honest conversation that is disciplined by an ethics of discussion. A conversation where the best discourses—the most persuasive ones—triumph spontaneously on the domination-free "market of economic ideas" (cf. Maki 1995). No market of ideas here but a field of struggle and cooperation between different epistemic communities, a hierarchical and institutional field of the production and evaluation of knowledge (Bourdieu 1975; Chavance and Labrousse 2018). These epistemic communities are also discursive communities, cemented by shared cultural and epistemological convictions (Beacco and Moirand 1995). Their discourse refers to other discourses held by other communities: they are conceived of "as an attack, a defense, a criticism or a contribution to a position or a particular set of thoughts." (Skinner 2003: 100).

[3] Parenthetical page indications refer to Banerjee and Duflo (2011).

This analysis is based on *Rhetoric* by Aristotle. It defines rhetoric as "the faculty to consider […] the available means of persuasion" (book 1, chapter 2). The Aristotelean core notions of *logos* (argumentative processes using reason) and *ethos* ("in which light does the speaker appear" (book 2, chapter 1): the qualities the speaker demonstrates through its speech) are the main organizing principles of this analysis. It also examines the "textual layout, conceived of as the sequencing of persuasive strategies within the text: the tactic of textual arrangement [and the] strategies for text expression (typodisposition, distribution, typography, punctuation, etc.)" (Duteil-Mougel 2005: 3). These are "evaluative devices": they provide indications as to what the author intends to highlight (Strassman and Polanyi 1995).

Here the epistemic effects of a discursive device are as important as its persuasive effects: in what way do its argumentation methods frame and shape knowledge, how do they allow certain phenomena to be conceived of and do they shroud other phenomena? This is an important aspect of discourse analysis: "discourses emerge as particular ways of construing (representing, interpreting) particular aspects of the social process that become relatively recurrent and enduring and which necessarily simplify and condense complex realities, include certain aspects of them but not others, and focalize certain aspects whilst marginalizing others" (Chouliaraki and Fairclough 2010: 1215). These processes of focusing, reducing, marginalizing, and removing from the scope, deserve particular attention.

8.1.2 Methodology and Outline

In this perspective, I began with a first linear reading of the book. This reading allowed me to identify the salient rhetorical processes relative to statistical, graphic and textual content in *Poor Economics.* Following this phase of abductive exploration, I proceeded to counting and the inductive inventory of occurrences (and in some cases co-occurrences[4]) of these elements. This count makes it possible to systematically examine the relative importance of terms and to observe variations of form and content in context.

I then tested the presence of other lemmas in the book in order to examine, in a deductive way, the solidity of the first results of the analysis, so as to highlight the out-of-discourse (the absent or sparse terms). If a term is very present then another related term is likely to also be very present; conversely, a conflicting term will likely be sparse. I also compared some aspects of *Poor Economics* to other popular science books by development economists (Easterly 2001; Sachs 2005; Stiglitz 2006) to pinpoint its rhetorical specificities.

[4] When the inventory revealed repeating associations, I quantified them.

The analysis of the rhetoric in *Poor Economics* begins with the mobilization of numbers (1), which is followed by the study of the use of graphic design (2), next, an examination of the presence and multiple rhetorical functions of the anecdotes in the book (3), and lastly the identification of two transversal and impactful narrative schemes (4).

8.2 Hard Numbers: The Rhetoric of Numbers, the Number as Rhetorical Figure

Numbers hold a special place in *Poor Economics*. Very present, they are accompanied by a rhetoric of the convincing number. Some figures are staged as arbiters of economic controversies and as representations of the lives of the poor.

8.2.1 Quantify and Disqualify

Poor Economics develops a rhetoric of proof by numbers, one that is found in all J-PAL productions. Here, the semantic field of experimental proof and numbers is significant. The book contains 130 instances of the term "evidence" (4 for "proof"), 102 times for "fact(s)" (+31 for *in fact*), 85 for "number,"* 84 for "experiment,"* 72 for "random,"* 57 for "control,"* 45 for "data," 18 for "trial,"* 9 for "RCT." By way of comparison, because they relate directly to its *leitmotiv*, the most significant terms in the book relate to the lives of the poor, this accounts for 584 instances for "poor,"* 207 for health-related terms ("health"*), and 150 cumulative instances for "life"* and "lives"* (including "lifetime" or "livestock").

The numbers themselves are ubiquitous: there are a total of 3607 figures in the development of the text.[5] These 3607 instances incorporate a total of 236 dates, highly concentrated in recent years (1990–2011). Compared to the corresponding number of pages, this makes an average of 12.9 digits per page (dates included) and 12.0 digits per page (excluding dates). Randomists seem to obey "Kelvin's Dictum: When you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind," which is the critical rule of economists' scientific credo, according to McCloskey (1983: 484). There are, therefore an average of 12 digits per page, even though the book contains no statistical table and 6 out of 7 graphs correspond to a theoretical representation that is not based on statistical data (see next section). Numbers are therefore mobilized in a "literary" narrative, a relatively unexpected feature. We will return to this idea later.

[5] This count excludes the table of contents, endnotes, acknowledgements, the presentation of authors and editor, as well as the page numbers, section numbers, and chapter numbers and references to them.

In J-PAL's rhetoric, only the figures from randomized controlled trials (which are qualified as a "new powerful tool" (p. 13)) are conclusive. "The studies we use have in common a high level of scientific rigor, openness to accepting the verdict of the data […]" (p. 16). Their methodology appears all the more rigorous as it is based on medicine: "the cleanest way to answer such questions is to mimic the randomized trials that are used in medicine to evaluate the effectiveness of new drugs" (p. 9). These figures are synonymous with objectivity. The rest is nothing but ideology and ignorance, evils that *hard numbers* provide the ability to combat. With inertia, these are the "three I's" that must be remedied in order to efficiently help those living in poverty. The authors expressly make it the central message of the book:

> The message of this book […is that] ideology, ignorance, and inertia—the three I's—on the part of the expert, the aid worker, or the local policy maker, often explain why policies fail and why aid does not have the effect it should (p. 16)

The easy to remember formula of the "3I's problem" appears five times in the book.[6] The RCTs are placed at the top of the evidence hierarchy while other methods are disqualified as ideological and inconclusive. In previous J-PAL productions, RCTs were presented as the *gold standard* (zero occurrence in the book), revealing a "methodological proselytism" (Jatteau 2016). Here, RCTs are simply the "cleanest way" to proceed. In 2007 Banerjee referred to international regressions and case studies as "wishy-washy evidence" (Labrousse 2010). This idea is more subdued in *Poor Economics,* first of all in inaugural developments that criticize macroeconomic regressions (pp. 3–5), the results of which are presented as uncertain and falling under "big philosophical questions" (p. 4), "speculating on the grand scale" (p. 5) and in the following paragraph on microfinance:

> Unfortunately, […] until very recently, there was in fact very little evidence, either way, on these questions. What CGAP calls evidence turns out to be case studies […]. (p. 167)

In a few passages from the book, numbers seem to speak for themselves: "The data squarely rejected this view" (p. 124), "The data seems to squarely hand victory to the demand wallahs" (p. 112), "[…] accepting the verdict of the data" (p. 16). There is a variation with "evidence," "Whose story—the activists' or the skeptics'—does the evidence support?" (p. 44). "The evidence suggests the opposite." (p. 50). "Our evidence shows" (p. 171). This figure of speech, in which

[6] These 3I's are also against-formula. They aim to replace the 3I's of their adversaries, mainstream institutional political economists. These 3I's (interests, institutions, and ideas) were put forward by political scientists and repeated by Acemoglu and Robinson.

data speaks for itself, is a "hypostasis": a fictive entity (data) is considered an active subject (Lalande 1902–1923). The decisive role of the randomist in experimental construction and the interpretation of these results, is therefore obscured. This points to Esther Duflo's more general point of view: "Evaluations are rigorous. They leave no room for interpretation. If it doesn't work, it doesn't work. The only thing left, then, is to try something else" (in: Labrousse 2016: 289–91).

8.2.2 Ninety-nine Cents, Synecdoche for the Life of the Poor

In *Poor Economics,* one number is repeatedly placed center stage: 99 cents. Representing the international threshold of absolute poverty, this figure appears 18 times in the book while its technical counterpart "poverty line" appears six times (of which four are in an explanatory note, p. 277). This is a synecdoche for the lives of the poor, systematically occurring simultaneously with "live/living": ("living on less than 99 cents a day/per day"). The first version of the "companion site" for *Poor Economics,* mentioned three times, was originally titled www.99centsthebook.com, an extra clue as to the importance of this number. It represents the life of people living in poverty, the subject of the book, and becomes a metonymy of the book itself.

Why 99 cents? Banerjee and Duflo (2011: 277, endnote) justify this choice of number by referring to the work of Deaton and Dupriez (data collected in 2005 for the International Comparison Program at the World Bank). The threshold for absolute (monetary) poverty is measured at 16 Indian rupees based on a basket of goods consumed by people in poverty. This is the equivalent of 99 cents, equal to its purchasing power (PPP) in the United States, adjusted for price indexes. Nevertheless, other numbers were available. In 1985, Ravallion popularized the number $1.02 in PPP (the basis of the famous phrase, "a dollar a day"), in 2008 the World Bank reassessed the threshold of poverty at $1.25 in the PPP of 2005. These thresholds are a source of controversy (Reddy and Lahoty 2016) and, like all data, rely on social conventions (Desrosières 1998; Porter 1995).

This number uses both *logos* (the previous "technical" argument) and *pathos*. Gripping, it makes poor people's situation tangible for the reader, who is necessarily rich (being able to buy the digital version of the book for $9.99). In the global North, 99 cents is presented as a "symbolic price," nothing at all for buying a small thing. In the global South, 99 cents is the maximum amount that the poor have at their disposal each day. For the American reader, it is an easily memorable and emblematic number of the consumer society, from which the poor is excluded: this palindrome number echoes the number after the punctuation in a supermarket price tag and the name of a US discount chain ("99 cents stores").

The highlighting of "hard numbers" allowing to escaping sterile alternatives, to combatting ideologies and to reaching purely objective truths, contributes to a depoliticized vision of experimental protocol and of international aid (Labrousse 2016). This "faith in numbers" (Porter 1995) is particularly prominent in mainstream economics. It echoes the way international organizations and development experts (Ferguson 1990) promote "'consensus building' techniques, which disqualify oppositions and conflicts and evade power relations" (Hibou 2011: 136).

8.3 Graphic Representations: Embodied and Metaphorical Storytelling, Cognitive Framing

The presence of graphs—seven in total—is relatively low in *Poor Economics*. This is surprising, considering the rhetoric of numbers and the omnipresence of graphs in the contemporary media (Koetsenruijter 2017). However, the graphs in the book do not present, with one exception, statistics on the lives of the poor, or the results of RCTs: rather, they are abstract economic representations. Testifying to the seriousness of the authors (*ethos*), these formalizations would likely alienate the uninformed reader if they were not accompanied by attractive narratives that feature real characters.

8.3.1 What Is Kennedy's World? Representing and Reducing the Realm of Possibilities to Two Diagrams

If used sparsely, these diagrams play a significant role. Here, the two first graphs (pp. 12–13) condense the main issue and backbone of the book: "are there poverty traps or not?" The importance of the idea of poverty traps is highlighted by its 46 instances in the body of the text. This is considerable for a specialized term: the more generic term "development" only appears 28 times. These graphs present two worldviews: the world according to Sachs (who believes in the existence of poverty traps and thus sees the world according to Figure 8.1) and the world of Easterly ("according to Easterly, there are no such things as poverty traps"), represented by Figure 8.2.

The two diagrams are accompanied by a commentary that turns beliefs in poverty traps into a question of faith. "For those who believe in poverty traps, the world looks like Figure [8.1]"; "Many economists (a majority perhaps) believe, however, that the world looks like Figure [8.2]." This question is correlated to a real-world character in the book, Kennedy: "So which of these diagrams best represents the world of Kennedy, the young Kenyan farmer?"

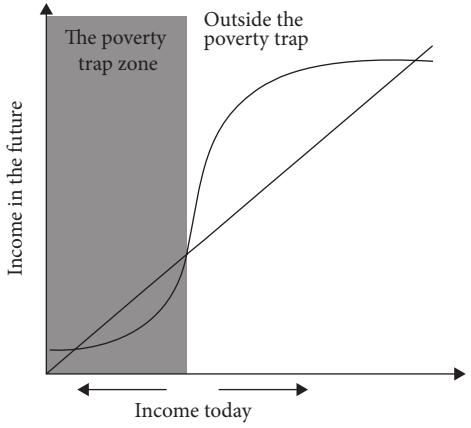

Figure 8.1 The S-shape curve and the poverty trap

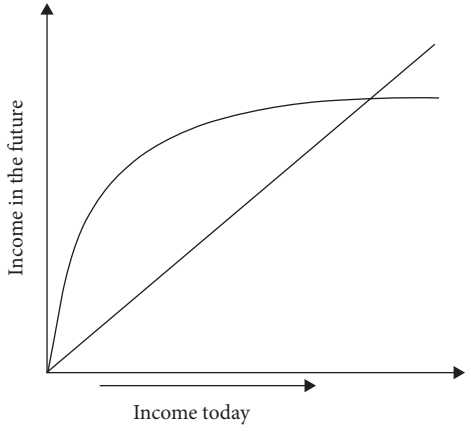

Figure 8.2 The inverted L-shape: no poverty trap

(p. 12). This question can only be answered with empirical (and not theological) evidence from RCTs:

> *we will find them in some areas, but not in others*. […] We will see many instances in the chapters that follow where the wrong policy was chosen, not out of bad intentions or corruption, but simply because the *policy makers had the wrong model of the world in mind*: They *thought there was a poverty trap somewhere and there was not, or they were ignoring another one that was right in front of them*.
> (p. 15, author's emphasis)

In Breton's typology, this graphic narrative falls into the category of *framing statements* and, more specifically, manipulative framing statements. Indeed, it rests

upon a "framework of false alternatives." The two alternative world views that are presented here act as blinders and considerably restrict the possible worlds. They leave out of the field, *without specifying this*—it is in this way that the framework is manipulative—fundamental questions of development economics.

These off-camera elements are, in fact, out-of-discourse. The organization of production and business, innovation dynamics, meso-economic and territorial questions, local and international financial and commodity flows, macroeconomic dynamics and politics, the environment and inequalities are largely absent. As such, there are no instances in the body of the text for *inequal** and *unequal*, *Gini coefficient*, *income/wage disparity/ies*, *justice*, *ethics**, *dependency*, *terms of trade*, *import*, *comparative advantage*, *commodity/ies*, *stabilization*, *specialization*, *international relation**, *industrial revolution, capitalism, market economy, modernization, westernization, globalization*, *tariff**, *reserves, foreign investment, capital flow*/flight**, *brain-drain, volatility*, *instability*, *speculation/tive*, *deregulation*, *Dutch disease, monetary policy, fiscal/budgetary policy, redistribution**, *protectionist**, *lost decade, (Post)Washington consensus, IMF, structural adjustment, foreign debt, foreign investment, fair/free trade, regional development, value chain, production network, corporate governance/interests, innovation fund, technology gap, patent, license, intellectual property, agrarian reform, land grabbing, deforestation, commons/common pool, natural resources*, *climate change, greenhouse (gas), biodiversity, public good*. *Industrial policy* appears only once, which is the same for *domination, dynamics (familial), inequity* (intrafamilial), *trade* (the idea of *trade credit*), *remittance, diversification* (of risks), *pollution* ("pollution inspectors"), *externalities* ("treatment externalities"), *global warming, carbon emission, liberalization ("early years of Chinese liberalization"), privatization* ("privatization voucher" for school fees) or *recession*. *Energy* is used only in the psychological sense (3 uses); the same is true for 5 out of 7 instances of *depression.* The results were similar for *structure* and *macro* (cf. Section 8.5). This is revelatory of the fundamental difficulty of RCTs in tackling historical dynamics (including microeconomic dynamics), and meso and macro questions. These issues are not amenable to RCTs.

These omissions are made apparent when comparing this book with other texts. First, texts from founders of development economics (Meier and Seers 1984), as well as reports from international institutions that have punctuated debates on world poverty since the 1980s: the UNICEF report (Cornia, Jolly, and Stewart 1987) on human damage of structural adjustment programs (SAP), the report from UNDP (1999) referring to "the lost decades" for Africa and Latin America, the World Bank report (2005) on the lessons from the failures of the PAS and shock therapies, writings from established economists on the *Post-Washington Consensus* (Stiglitz 2004; Rodrick 2008). It should be noted that the aforementioned keywords, with the exception of *land grabbing, land reform, production network*, and *dependency*—are all found in the popular science book *Making Globalization Work*

by Stiglitz (2006). These "blanks" of the RCTs also surface when compared with notorious economic literature on inequalities—from Bourguignon to Piketty[7]—or with heterodox political economy. Nevertheless, the 2015 World Bank report "Mind, Society and Behavior" overall lines up with Banerjee and Duflo's point of view (World Bank 2015). In this point of view only the microeconomic scale is rigorous, thus ignoring the meso-economic level and making macroeconomics a field of semi-metaphysical speculation.

8.3.2 The Extended Metaphor of the S-curve: When Ibu Tina Fell into the Poverty Trap

All of the graphs in *Poor Economics* are constructed around the following question: does the S-curve of poverty trap exist or not? In particular, this is the case for the only graph based on statistical data (Wealth in 1999 and 2005 in Thailand, p. 201). Nothing is said of the reasoning behind choosing Thailand or this particular time period (1999–2005), while at the same time the commentary is very general. Contrary to other graphs, it is unclear and little is clarified by the authors. It is, in some way, an "*ex machina* graph." Its sudden and mysterious intervention has the advantage of revealing an S-curve, aligning it with the real world phenomena. It is introduced with an emphatic "do" and a touch of humor (torturing the S):

> *We do see this S-shape between net worth today and net worth in the future in the real world*. [The graph] plots the relationship between resources the households had in 1999 and what they had five years later in Thailand. *The curve has a flat, elongated S-shape* (admittedly, we are torturing the S a little bit). […] What is more distinctive is the way in which the relation is fairly flat at very low levels of resources but then turns up sharply before flattening off. *This S-shape, as we saw before, generates a poverty trap*. (p. 200, author's emphasis)

Another key graph, "The impact of shock on Ibu Tina's wealth" has an S-curve. It appears earlier in the book (p. 139). It prepares a very general story around the S-curve and makes it tangible. This graph shows another real-world example of a person, Ibu Tina, tipping into a poverty trap. It is preceded by the story (393 words) of her life. It tells a tragic event, which caused this person's life situation to be reversed: a robbery quickly pushes Ibu Tina and her family into poverty, from which she will be unable to escape. This text is followed by its graphical representation: "In Figure [8.4], we have plotted the relationship between income today and income in the future for Ibu Tina, the Indonesian businesswoman." (p. 139).

[7] Piketty, who played an important role in Duflo's arrival at MIT, is cited in the acknowledgements but not in the rest of the book.

This graph is then translated into a conceptual commentary. It is the subject of a more pronounced economic vocabulary and an economic morality (the last sentence of the following excerpt):

> Before the debacle of the bounced check, Ibu Tina and her husband were outside the poverty-trap zone. If we follow their path over time, we see that they were on the trajectory to eventually arrive at a decent income. But the theft wiped out all their assets. This had the effect of moving them to the poverty-trap zone. Thereafter, they made so little money that they kept getting poorer over time […] When the relationship between income today and income tomorrow is S-shaped, a family can plunge from being on a path to middle class to being permanently poor. (p. 139)

This graph allows the reader to *visualize* how Ibu Tina literally falls into a poverty trap of which she will remain a prisoner. This fall changes the course of her happy destiny which was moving towards a reasonable income. This is an original method of graphic narration that combines *logos* (here the abstract character of the graph) with an individual embodiment (Ibu Tina) and a singular event (the theft) related to *pathos,* part of an expressive narrative that is both graphic and verbal.

These S-curves are a structuring and recurring metaphor (no less than 30 instances for S-Shape). It portrays poverty traps and is repeated in the other graphs. This metaphor is presented as a lived reality or a shared and active belief among economic actors, like teachers, students' parents or this shopkeeper from Gulbarga that the authors met:

> As we saw, *the belief in the S-shape curve leads people to give up*. If the teachers and the parents do not believe that the child *can cross the hump and get into the steep-part of the S-curve*, they may as well not try: The teacher ignores the children who have fallen behind and the parent stops taking interest in their education. (p. 91)
>
> […] once a micro-entrepreneur *realizes* that she is probably *stuck in the low part of the S-curve* and will never be able to make that much money, it may [be] difficult for her to be fully committed to her business. Imagine an entrepreneur who is below point M in Figure 3. It could be the shopkeeper we met in Gulbarga. (p. 222, author's emphasis)

This S-shaped curve, emerging from the minds of economists, would thus be part of the lived world of the poor and not the authors' interpretative schemes. It becomes all the more expressive as the metaphor is rolled out: in the school obstacle course, the impoverished students are afraid of crossing the hump or

gaining on the steep part of the curve, causing them to fall behind academically; the micro-entrepreneur realizes that she is *stuck* in the lower part of the S-curve. The bottom of the curve is the realm of the poor.

> *The shape of the curve is key*: It is very flat at the beginning, and then rises rapidly, before flattening out again. We will call it, with some apologies to the English alphabet, the S-shape curve. *The S-shape of this curve is the source of the poverty trap*. (p. 10)

This last sentence appears twice with a slight variation: "The sinusoidal character of the curve is at the origin of / generates the poverty trap" (10 and 200). Surprisingly, these sentences literally describe the shape of the curve as the cause of the trap, although it would be assumed that the S-curve "represents" the poverty trap. It seems to be magical thinking. Closely related to this curve, the idea of the poverty trap is also a metaphor that touches on a neighboring semantic field. The individual can "plunge" into these traps and remain trapped (12 instances for "trapped," six "trapped in poverty"), or "stuck" (p. 43). This metaphor is elaborated with ladders that allow traps to be escaped:

> As he [Jeffrey Sachs] sees it, there are healthbased *poverty traps*, but there are also *ladders* we can give to the poor to help them *escape from these traps*. If the poor cannot afford these *ladders*, the rest of us should help them out. (p. 46)
>
> The *ladders to get out of the poverty trap* exist but are *not always in the right place*, and people do not seem to know *how to step onto them* or even want to do so. (p .50)

It is also a matter of releasing the trap ("to set the trap loose," p. 42, to "break the trap" (p. 234), to "escape the trap" pp. 10 and 46 and "get out of the trap" (title p. 200). As McCloskey (1983: 502) has demonstrated, "economics is heavily metaphorical" and its "models are metaphors." *Poor Economics* epitomizes this, pursuing the art of metaphor and anecdote to an exceptional extend.

8.4 A Staggering Wealth of Anecdotes

The story of Ibu Tina is just one anecdote among many others. However, when examining the 11 instances of the concept of anecdotes (or the adjective "anecdotal") in the text, a body of doctrine that is hostile towards anecdotes emerges (Inset 1).

8.4.1 Doctrine: RCT Data, Antidotes for Misleading Anecdotes?

Inset 1: Inventory of instances of anecdotes in *Poor Economics*

1. "If the poor appear at all, it is usually as the *dramatis personae of some uplifting anecdote* or *tragic episode*, to be *admired* or *pitied*" (p. viii)
2. "We *need evidence*. But unfortunately, the kind of data usually used to answer the big questions does not inspire confidence. There is *never a shortage of compelling anecdotes*, and it is *always possible to find at least one to support any position*" (p. 4)
3. "However, there are now *a number of careful experiments that suggest that such anecdotes are oversold*." (p. 57)
4. "For all the *individual anecdotes* of fruit sellers turning into fruit magnates that can be found on the various Web sites of microfinance institutions, *there are still many poor fruit sellers in Chennai*" (p. 159).
5. "We met a prominent Silicon Valley venture capitalist and investor, and supporter of microcredit […], who told us that he *needed no more evidence. He had seen enough*
6. *"anecdotal data" to know the truth.*"
7. "*Anecdotal data does not help with the skeptics out there,* including large sections of governments everywhere […]."
8. "*The anecdotes* […] *did little to help them out.* One reason the MFIs were lacking a *powerful argument* in their defense is that they had been *reluctant to gather rigorous evidence to prove their impact.* […]"
9. "*Anecdotal data is not truth or evidence.*" (p. 168).
10. "The MFIs *responded to the evidence from the two [RCT] studies* […] *with six anecdotes* on successful borrowers" (p. 172).
11. "A *study* […] shows that this role […] *goes beyond this particular anecdote*" (p. 228).

In *Poor Economics,* anecdotes are openly and fundamentally misleading. They are presented as deterrents, able to be manipulated at will (*it is always possible to find at least one [anecdote] to support any position*), enabling certain positions to be oversold, arousing various emotions: fascination (*compelling*), comfort (*upflifting*), admiration (*admired*) or sympathy (*pitied*), but not reason. Anecdotes are also systematically put in opposition with the realm of evidence (*evidence, to prove*) and truth stemming from RCTs (*experiments, studies*), which are associated with rigor (*rigorous)* and meticulousness (*careful*). At best, anecdotes—classically in academic literature—stem from particular cases that cannot be generalized (the few individuals who became fruit magnates aren't the majority of fruit sellers). Yet, some actors and micro-finance institutions are "reluctant to gather rigorous evidence." They thus misuse anecdotes that they have mistaken for proof. Fortunately, the moral of the story (on p. 168) is the following: for those who do not listen to the sirens of "anecdotal data," the anecdote is of little help, feeds into a well-founded skepticism and ultimately, "lacks a powerful argument." Only RCTs allow this: "demonstrations of the persuasive power of a successful randomized experiment" (p. 78) are mentioned. Anecdotal data is a decoy.

8.4.2 The Practice of Anecdotes: A Paradoxical Plethora

McCloskey (1983: 482) reported a phenomenon that seems to apply to anecdotes in *Poor Economics*:

> ECONOMISTS DO NOT FOLLOW the laws of enquiry their methodologies lay down. […] Their genuine, workaday rhetoric, the way they argue inside their heads or their seminar rooms, diverges from the official rhetoric.

While numbers are ubiquitous in the book, they are not mobilized as antidotes to anecdotes. On the contrary, they function in symbiosis with the anecdotes. Looking at the structure of the book, many chapters open with an anecdote, typically mentioning the name of the person whose story is being told. For example, the title of chapter 5 is associated with the name of the main protagonist of one anecdote: "Pak Sudarno's Big Family." Anecdotes, further, are not just an introduction but are also found throughout the developments of each chapter. Their systematic presence is not anecdotal. The most common and striking stories are about the poor (see Inset 2). These are the "*dramatis personae*" of the book.

Inset 2 Named after people: 12 anecdotes which set the stage for the poor

1. The story of *Pak Solhin* (Indonesia): unemployed farm worker, occasional fisherman, nutritive poverty trap, 16 instances
2. The story of *Allal Ben Sedan* (Morocco): small breeder who did not want microcredit, 13 instances
3. The story of *Xu Aihua* (China): poor person who became a successful entrepreneur, 12 instances
4. The story of *Pak Sudarno* (Indonesia): scavenger, father of nine children, demographics, 10 instances
5. The story of *Michael and Anna Modimba* (Kenya): corn farmers, access to fertilizers, 10 instances
6. The story of *Kennedy* (Kenya): farmer, use of fertilizers, 9 instances
7. The story of *Jennifer Auma* (Kenya): sells grains and legumes at the market, savings for the poor and tontines, 8 instances
8. The story of *Ibu Emptat* (Indonesia): weaver's wife, health-related poverty trap, 7 instances
9. The story of *Shantarama* (India): widow and mother of six children, school absenteeism, 5 instances
10. The story of *Wycliffe Otieno* (Kenya): farmer, access to fertilizers, 5 instances
11. The story of *Pak Awan* (Indonesia): unemployed construction worker, opening up a shop for lack of better options, forced entrepreneurship, 4 instances
12. The story of *Oucha Mbarbk* (Morocco): occasional construction and agricultural worker, preference for entertainment, starves himself to buy a TV, 2 instances

---

8.4.3 Anecdotes Following Social Marketing and Storytelling Principles

These real-world people embody an emblematic problem in the economics of poverty. For example, Pak Sohlin exemplifies the poverty trap related to a shortage of food: "Pak Solhin […] once explained to us exactly how such a poverty trap worked" (p. 23). His story is detailed in 577 words covering a span of three pages. Some of these poor are reccurring characters: their story is told in one chapter and they come back in other episodes (16 instances for Pak Solhin). These stories meet the principles of storytelling: they are personal, true, simple, informative, concrete, detailed, emotional and actionable (Few 2009). In the words of Banerjee and Duflo: "the point is simple: talking about the problems of the world without talking about some accessible solutions is the way to paralysis rather than progress" (p. 5). The goal of this storytelling is to be persuasive as to the utility of experiments and the solutions that they verify, to appeal to donors. This storytelling participates in social marketing.

8.4.4 Anecdotes and Charitable Efficiency: An Effect Demonstrated by Experiments

Banerjee and Duflo, through the programs they evaluate, are familiar with social marketing techniques. These techniques have been applied to development aid and health policies in the global South, particularly in India. One such component is the paternalistic nudge that the authors assert. The impact of social marketing on charitable giving is a major research area for RCTs in the US (List and Lucking-Reiley 2002).

Starting on the second page of the book—a clue as to the importance of the subject—Banerjee and Duflo mention the results of an experiment on the efficiency of two types of donation requests for a charitable organization. In the first group, the donation requests use scientific vocabulary (pure *logos*) and are based on abstract data on food shortage as a result of low rainfall. In the second group, which raises twice as many funds, the donation requests are personalized and emotional (*pathos*). They feature the little Rokya, seven years old and "desperately poor and threatened by hunger," whose life can be changed thanks to donations. The use of personalized anecdotes by Banerjee and Duflo here finds one of its raisons d'être; there are others.

8.4.5 Anecdotes Testifying to the Authors' Ethos: Proximity, Familiarity, Credibility

The anecdotes of *Poor Economics* are personal for two reasons. First, they are personalized by the first and last name of the character. Second, they are personal because they are almost always people that Banerjee and Duflo met in the field. These anecdotes thus speak to the authors' proximity to the poor, an essential component of their *ethos* in the book. Starting in the preface, the authors present themselves as the modest guests of the poor to whom they are thankful (p. viii). The testimony of their time spent with the poor is a *leitmotiv:* there are 34 instances of "[name] we [the authors] met." This experience from the field is highlighted, here again, in the preface:

> We are academics, and like most academics we formulate theories and stare at data. But the nature of the work we do has meant that we have *also spent months, spread over many years, on the ground* working with NGO activists and government bureaucrats, health workers and microlenders. This has taken us to the *back alleys and villages where the poor live, asking questions, looking for data*. (p. vii)

It is a source of "comparative rhetorical advantage" for the authors against rivals like Acemoglu and Robinson (see Section 8.5). Few *mainstream* economists can boast as much "field" work as the authors of *Poor Economics*. The profession operates first indoors, through remote data processing. However, from the perspective of other social sciences in which there is more demanding approach to fieldwork, the field in *Poor Economics* appears too short and not rigorous enough (see Jatteau, 2013 and Section 8.6).

Additionally, there are also anecdotes regarding the authors themselves. Both authors are identified by their first name only, Esther (47 instances) and Abhijit (38 instances). This is the case right from the first lines of the book:

> Esther was six when she read in a comic book on mother Theresa that the city then called Calcutta was so crowded that each person had only 10 square feet to live in. She had a vision of a vast checkerboard of a city, with 10-foot squares marked out on the ground, each with a human pawn, as it were, huddled into it. She wondered what she could do about it. […]
>
> At six, Abhijit knew where the poor lived. They lived in little ramshackle houses behind his home in Calcutta. Their children always seemed to have lots of time to play, and they could beat him at any sport: When he went down to play marbles with them, the marbles would always end up in the pockets of their ragged shorts. He was jealous. (p. vii)

Such a process engages in *captatio benevolentiae* (Dokova 2016), a classic oratory process aimed at attracting the benevolence and sympathy of an audience from the exordium (beginning of speech). It creates *familiarity with the audience*. This familiarity is strengthened here through the way that the authors are presented as naïve young children (a pledge of modesty), and who are subject to very human passions like jealousy. This process is recurring:

> When a puzzled six-year-old Abhijit asked his Bengali aunt […] (p. 70)
>
> Abhijit was falling behind in his schoolwork in first grade […] (p. 90)
>
> The school Abhijit went to in Calcutta […] (p. 94)

This familiarity is fostered by references to other members of Abhijit's family. In addition to his aunt, his father, mother and grandfather are also the subjects of anecdotes (pp. 70 and 183).

The authors' *ethos* is completed with their academic presentation (pre-discursive ethos[8]). The familiarity induced above is combined with multiple pledges of credibility and scientific authority. This presentation details the authors' *cursus honorum* and their exceptional academic and symbolic capital (the most prestigious diplomas and awards). In the rest of the book, Esther and Abhijit are also represented as adults, experienced scientists and experimentalists:

> But several simultaneous experiments that Esther, Pascaline Dupas, and Michael Kremer conducted (p. 114)
>
> To evaluate it, Esther compared (p. 81)
>
> In a study with Udry, Esther found (p. 125)
>
> Abhijit with two Chinese-born co-authors, Nancy Qian […] and Xin Meng (p. 120)

Here, co-authors' last names are mentioned, contrary to simply "Abhijit" and "Esther." Laid end-to-end, these personal anecdotes about the poor, as well as about the authors and their relatives, illustrate a remarkable portrait of the authors' ethos. They are benevolent towards poor people and familiar with their situation, they are modest and accessible even though they have accomplished excellence in their careers. This guarantees the authority of their statements.

[8] Annexed from discourse: formally, it is not the authors who present themselves. "The pre-discursive ethos refers to the reputation of the speaker […They] are at work in a system based on auctoritas" Duteil-Mougel (2005: 4).

8.4.6 Anecdotes, Didactics, and Distinction

Anecdotes have long been part of the pedagogical arsenal (Stock 1993, Ford 2002). Among the "examplification processes," anecdotes fulfill a didactic function: "a real or simulated intention […] to bring new knowledge to others" (Beacco and Moirand 1995: 33). These short stories give flesh to the data, allowing the authors to attract and keep readers' attention throughout a relatively long book. The J-PAL members, because they regularly address non-academic audiences (heads of NGOs, international organizations, policy-makers, etc.), have developed extensive didactic abilities. Kuhn (1962: 187) demonstrated the centrality of *exempla* in learning a paradigm and in the differentiation between scientific communities: "More than other sorts of components of the disciplinary matrix, differences between sets of exemplars provide the community fine-structure of science." In *Rhetoric* (book 2, chapter XX), Aristotle distinguishes groups of examples involving historical facts from fabricated examples. Here, examples are true stories and not fables (the tale of bartering and other Robinsonades) or imaginary examples, a narrative method that is very common in economic textbooks (Jallais 2018).

The authentic anecdotes of *Poor Economics*, those which touch upon the lives of impoverished people above all, thus distinguish the epistemic community of the J-PAL from other communities in the disciplinary field. Anecdotes also distinguish economists in the field of "practitioners of development." *Poor Economics* can be compared to other popular science works written by economists who, due to their roles in international organizations, have completed multiple trips in the field. For example, in Stiglitz (2006), there are very few anecdotes. Sachs (2005) largely uses personal anecdotes in *The End of Poverty.* However, these anecdotes mainly relate to *president* (70 instances), *minister* (50), *leader* (106), *secretary-general* (17) and others, like *director* (14). This is symptomatic of his "top-down" vision as an advisor to the prince. The poor met in the *Millennium Villages* are more marginally the—anonymous—subjects of the story. Easterly (2001), in *The Elusive Quest for Growth* mentions some anecdotes about the poor, especially in "intermezzo" and "Leila story." Some of these stories stem from his trips in the field, while others are from the media. Yet, Easterly's main stories relate to entire countries (Ghana, Ivory Coast, India, China, etc.). These big stories structure the book, a similarity with Sachs (stories about Bolivia, Poland, Russia, China, etc.) and a contrast with Banerjee and Duflo, who focus exclusively on "small," personal stories. Unlike the latter, none of these economists develop a hostile doctrine towards anecdotes: there is no instance of this notion in Sachs, one in Stiglitz, five in Easterly, though none with negative connotation.

8.4.7 The Discreet but Inchoate Heuristic Function of Anecdotes

In Banerjee and Duflo, the role of anecdotes is not limited to ethical, rhetorical, and pedagogical dimensions. In practice, anecdotes play a hidden, though crucial heuristic role. In a radio interview with France Culture (January 6, 2012), Esther Duflo explained that these stories "make the statistical data more understandable." In *Poor Economics,* Banerjee and Duflo report that these stories allow them to "knit together a coherent story":

> Many stories were shared with us. Back in our offices, *remembering these stories and analyzing the data*, we were both fascinated and confused, struggling to fit what we were hearing and seeing into the simple models that […] professional development economists and policy makers use to think about the lives of the poor. […] This book comes out of that interchange; it represents our attempt to *knit together a coherent story* of how really poor people live their lives. (p. viii)

Resorting to stories in order to understand data illustrates the incompleteness of numbers. This is particularly true when it is a question of understanding the *why* behind surprising outcomes and when making sense of data. Indeed, as I have demonstrated elsewhere (Labrousse 2017): "It's tricky to grasp the causal path (How? Through what mechanisms?) that leads to a particular set of observed outcomes (whether it works or not). Indeed, aside from instances of simple mono-causality (a cause brings about an effect, with no feedback of the effect onto the cause), randomized experiments provide evidence of effectiveness (a particular effect is observed) rather than causality (what mechanisms generated this effect?)[9] […] In cases of complex, cumulative, multifactorial, and non-linear causality, causal chains becomes a kind of black box for experimenters."

However, the storytelling of *Poor Economics* is too thin and imbued with a priori assumptions to uncover sound causalities and, hence, often misleading. For instance, its light narratives stress the failings of the poor, like the tendency to spend rather than save, and to spend on the "wrong" things as ceremonial rituals, "wasteful" expenditures like buying tea (pp. 37, 171, 183–204). Ethnographic evidence from India points that teashops are a node for network-building and information gathering, that ceremonial gifts are a relational form of saving: "households, including those at the bottom of the pyramid, do save, in the sense of storing, accumulating, and circulating value. But this takes place via particular forms of mediation [like ceremonies] that allow savers to forge or maintain social

[9] On the difference between evidence of difference-making and evidence of mechanism (i.e. causality), see Berriet-Solliec et al. (2014).

and emotional relations, to keep control over value […] People prefer to create value – and save – by investing in their social networks rather than locking their assets in a bank account. […] Asserting that the poor lack self-control, a sense of time or the discipline to resist social pressure […] shows a total misunderstanding of local social and economic dynamics." (Guérin, Venkatasubramanian and Kumar 2019, 1 and 13). This alternative, thicker narrative relies notably on the careful observation of ceremonies and of notebooks where the poor keep a detailed accounting of contributions and receipts for each ceremony, and in-depth interviews about each transaction. In the *Proempleo* RCT, follow-up qualitative interviews avoided a complete misunderstanding of the RCT data on wage-subsidies (Ravallion, Chapter 1, this volume). The *Al Amana* RCT also exemplifies how ethnographic material challenged J-PAL narratives on micro-credit demand (Morvant-Roux et al. 2014). These "thin narratives" oppose the "thick description" of ethnographers (Geertz 1973).

8.5 Two Rhetorical Schemes with Strong Epistemic and Persuasive Effects

Far from this socio-economic critique, the rhetoric of *Poor Economics* is comfortably anchored in mainstream economics, the only type of economics referenced. This rhetoric provides a "persuasive advantage" in the competition between evidence within mainstream economics. It enables the book to combat explanations from competing schools of thought. Two transversal rhetorical schemes are particularly effective here: (1) The *rhetoric of the middle way between two extremes* reinforces the reasonable a-ideological posture of the J-PAL; thus enabling the discreditation of Ivy League^{+} competitors like Sachs and Easterly. This polarizes the debate and conceals the wide range of remaining approaches. (2) The *rhetoric of small measures with great effects* allows authors to inflate the micro and downplay the macro: it legitimizes the J-PAL's position, disqualifies political economy and the institutionalism of Acemoglu and Robinson.

8.5.1 The Middle Way between Two Extremes: Common Sense, Objectivity, and Manipulative Framing

Two influential economists will embody the two opposing poles of development economics: Jeffrey Sachs (39 occurrences) on one side and William Easterly (33 occurrences) on the other. They are antitheses, figures of symmetrical oppositions, of which Aristotle (Rhetoric, book III, chapter XIX) says: "Such a style is pleasing because opposites are most knowable and more knowable when put beside each other." In this example, Sachs and Easterly are described as each-having-a-universal-response-to-everything-yet-opposite-on-everything. Starting with the introduction,

their opposition is portrayed as a petty dispute between two neighborhoods in Manhattan:

> *Jeffrey Sachs*, adviser to the United Nations, director of the Earth Institute at *Columbia University in New York City* [..], *has an answer to all these questions*: […].
>
> But then there are others, *equally vocal, who believe that all of Sachs's answers are wrong. William Easterly*, *who battles Sachs from New York University at the other end of Manhattan*, has become one of the most influential anti-aid public figures […]. (p. 2)

According to the authors, these disagreements are ideological:

> It is *no accident that Sachs and Easterly have radically opposite views on whether bed nets should be sold or given away*. The positions that most rich country experts take on issues related to development aid or poverty tend to be *colored by their specific worldviews* […] *on the left of the political spectrum, Jeff Sachs* (along with the UN, the World Health Organization, and a good part of the aid establishment) wants to spend more on aid […]. *On the right, Easterly, along with Moyo, the American Enterprise Institute, and many others, oppose aid* […]. (pp. 8–9)

This ideological coloration discredits both opponents. In contrast, Banerjee and Duflo come across as being the voice of reason and common sense. It seems that the authors are not coming from an ideological standpoint or a principled stance, but rather from objectivity and empirical evidence. They address the problems concretely, one by one (see Inset 3).

Inset 3: Far from the ready-made answers of Sachs and Easterly, concrete responses

1. This book is an invitation to think again, again: to turn away from the feeling that the fight against poverty is *too overwhelming*, and to start to think of *the challenge as a set of concrete problems* that […] can be *solved one at a time*. (pp. 1–2)
2. There are in fact *answers*—indeed, this whole book is in the form of an *extended answer*—it is just that they are *not the kind of sweeping answers that Sachs and Easterly favor*. (p. 3)
3. This is why it is really helpful to think in terms of *concrete problems*, which can have *specific answers* (p. 5)
4. These questions can be *answered*, but the answers are by no means obvious. Yet *many "experts" take strong positions on them that have little to do with evidence*. (p. 6)
5. This radical shift in perspective, *away from the universal answers*, required us to *step out of the office and look more carefully at the world*. (p. 13)
6. there is *no general rule here* […]. It is the body of knowledge that grows out of *each specific answer* and the understanding that goes into *those answers* that give us the *best shot* at, one day, *ending poverty*. (p. 14)
7. […] although we have *no magic bullets* to eradicate poverty, *no one-shot cure-all*, we do *know* a number of things about how *to improve the lives of the poor*. (p. 267)

The authors thus come across as more meticulous, more realistic, and more modest: they have solutions, not miracle, universal[10] solutions but solutions that are adapted to each problem. They are involved in the concrete and not in speculation. This rhetorical scheme complements the authors' ethical portrait. It corresponds to a manipulative framing statement, a false alternative between two ideological extremes. Here, the authors omit all approaches that do not fit into these two forms of neoliberalism.[11] This is the case for Rodrik's or Stiglitz's approaches (never referenced), of the many trends in political economy that are not mainstream, such as classic development economics (Myrdal, Hirschman, Boserup, etc.).

8.5.2 Small Causes, Big Effects: Oxymorons in Defense of the "All Micro"

In the excerpt below, the previous scheme is combined with a second, structuring scheme, a rhetoric of "think small to solve global problems":

> We are often asked why we do what we do: "Why bother?" *These are the "small" questions*. *William Easterly*, for one, *criticized randomized control trials* (RCT) on his blog in these terms: "RCTs are infeasible for many of the *big questions* in development, like the *economy-wide* effects of good institutions or good *macroeconomic* policies." Then, he concluded that "embracing RCTs has led development researchers to *lower* their ambitions." This statement was a good reflection of an institutionalist view that has strong currency in development economics today. […] It follows (or so it is assumed) that *"big questions" require "big answers"—social revolutions*, such as a transition to effective democracy. *At the other extreme*, *Jeffrey Sachs* sees corruption, perhaps not surprisingly, as a poverty trap: Poverty causes corruption, and corruption causes poverty. (p. 236)

To contradict the rhetoric of "big questions—big answers" the authors will deploy the rhetoric of "small causes—big effects". This scheme is recurring. It appears as early as the preface:

> the *small* costs, the *small* barriers, and the *small* mistakes that most of us do not think twice about *loom large in the lives of those who have very little*.

[10] One can nevertheless question this point: "even if Duflo refuses to assess the helpfulness (or the harmfulness) of 'aid' or 'education' in general […] the locus of 'good' or 'bad' just shifts at a more concrete level of analysis: for the J-PAL some micro-social devices can be intrinsically good or bad for the poor in each given domain (education, nutrition etc.)." (Labrousse 2016: 286)

[11] Centralizing neoliberalism for Sachs and decentralizing neoliberalism for Easterly (in the Austrian tradition).

> It is not easy to escape from poverty, but a sense of possibility and a *little bit of well-targeted help* (a *piece* of information, a *little nudge*) can sometimes have *surprisingly large effects*.
>
> On the other hand, misplaced expectations, the lack of faith where it is needed, and *seemingly minor hurdles can be devastating*. A *push on the right lever* can make a *huge difference*, but it is often difficult to know where that lever is. Above all, it is clear that *no single lever will solve every problem*. (p. x)

It is found in eight instances of "small changes," including:

> *small changes*, we believe, can sometimes end in a *quiet revolution* (p. 237)
>
> What is not recognized as often, however, is *how important the effect* of *seemingly very small changes* can be. (p. 246)
>
> a *small change* in the rules *changed everything* (p. 249)
>
> don't let the *apparent modesty* of the enterprise fool you: *Small changes can have big effects* (p. 272)

It is repeated with the adjective "minor":

> *seemingly minor* interventions can make a *significant difference* (p. 253)
>
> *seemingly minor* hurdles can be *devastating* (p. x, preface)
>
> A *seemingly minor* technical fix, *involving no major political battle*, *changed the way* in which the voice of the poor was taken into account (p. 247)

There is also a variation with the adjective "incremental":

> We are *not "lowering* our ambitions": *Incremental progress* and the *accumulation of these small changes*, we believe, can sometimes *end in a quiet revolution* (p. 237)
>
> These *changes will be incremental*, but they will *sustain and build on themselves*. They can be the *start of a quiet revolution*. (p. 265)

The same idea is expressed with "step": "small steps" (1) and "step-by-step" (1), "first step" (9), "stepping stone" (2), which are opposed with "extreme steps" (1), "drastic steps" (1). These are found in the metaphor of bricks which, lead, one by one, to the construction of a house. "Brick by brick" is even the title of chapter 8. There are five instances counted of this phrase, and 14 of "brick."

This *leitmotiv* uses several figures of speech. Evocative metaphors such as baby steps, footsteps and other steps forward towards progress or bricks building a house. Above all, Banerjee and Duflo systematically use oxymorons: *quiet revolution, small/big, minor/significant* which reveal an apparent paradox (*seemingly).* The oxymoron allows authors to describe a situation in an unexpected and

previously inconceivable way. It is part of a strategy of surprise and pathos (Monte 2007). The contradiction in terms of these oxymorons is all the more apparent. Thanks to the almost systematic co-occurrence of the *seemingly* and *apparent* modalizations, the authors inhibit any judgment of contradiction in order to be more persuasive.

Another property of the oxymoron is to take the opposite position of doxa (Monte 2007). It is the doxa of macroeconomic and institutionalist approaches that must be discarded here. These are *oxymorons of combat*: according to Angrist and Pischke (2010), economists who are close to Banerjee and Duflo, empirical microeconomics is an assault on the "theoretical macroeconomic fortress." The brick metaphor also serves the promotion of this idea. Duflo previously justified the primacy of the micro over the macro while using the metaphor of the Meccano, a construction game:

> Using macroeconomic data […] leads to a stalemate. […] the macroeconomic model is constructed like a Meccano set, based on microeconomic building blocks […] In any case, the basic elements are microeconomic elements. (Duflo 2009: 73–4)

The macro is only the sum of micro behaviors, just as the house is only the sum of its bricks. This reductionism is typical of a *fallacy of composition* (Labrousse 2010 and 2016). Also covered are Acemoglu and Robinson as well as institutional political economics:

> Our *colleague Daron Acemoglu*, and his long-term coauthor, *Harvard's James Robinson*, are two of the most thoughtful exponents of the rather *melancholy view*, active in economics today, that until political institutions are fixed, countries cannot really develop, but institutions are hard to fix. Both *political scientists and economists* typically think of institutions at a very high level. They have in mind, if you like, *institutions in capital letters*—economic INSTITUTIONS like property rights, or tax systems; political INSTITUTIONS like democracy or autocracy, centralized or decentralized power, universal or limited suffrage. (pp. 236–7)

The repeated use of this striking contrast between key concepts, in upper-case (the macro, the abstract, the "broad"), and the lower-case (the micro, the concrete, the specific), comes to visualize and reinforce the process:

> To really understand the effect of institutions on the lives of the poor, what is needed is a shift in perspective from INSTITUTIONS in capital letters to institutions in lower case—the "view from below." (p. 243)

> The focus on the *broad INSTITUTIONS* as a necessary and sufficient condition for anything good to happen is somewhat *misplaced*. The political constraints are real, and they make it *difficult to find big solutions to big problems*. But there is considerable slack to improve *institutions* and policy at the *margin*. (p. 263)

There are five instances of this typographic contrast. The disclaimer "aid" rather than "Aid" is also used. Duflo and Banerjee thus go on a crusade "AGAINST POLITICAL ECONOMY" (the level 4 title) for which they provide a curious definition:[12]

> Political economy is the view (embraced, as we have seen, by a number of development scholars) that politics has primacy over economics: Institutions define and limit the scope of economic policy. (p. 252)

They mock the opposition between Sachs and Easterly's two ends of Manhattan. Here, they are the protagonists of a struggle within MIT's economics department (with Acemoglu) and with the government department of the neighboring Harvard (Robinson), the two being within a few square miles of each other in Cambridge (Mass.). The authors' goal is to persuade readers of their approach of "thinking small to fight global poverty" as the best possible approach. Persuasion is mentioned at the beginning of the introduction and in its last sentence, an indicator if its importance:

> The problem [of world poverty] seems *too big, too intractable*. *Our goal with this book* is to <u>persuade you</u> not to. (p. 1)
>
> We hope to <u>persuade you</u> that our *patient*, *step-by-step* approach is not only a *more effective way to fight poverty* [...]. (p. 15)

This scheme legitimizes the J-PAL's multiplication of experiments. Development is reduced to the implementation of a series of small devices aimed at influencing individual and group behavior. These *nudges* provide the impetus for modifying incentives (30 instances of *incentive**) and to steer impoverished people towards good behavior. This rhetoric of the "think small" eschews burning questions: increasing inequalities, the imbalance of international power relations, etc.[13] What is beyond the scope of the book is revealing. There are only five cumulative instances for *structure* or *structural.* Three of these instances concern

[12] For further discussion see Labrousse (2016).

[13] This distaste of broad-based policy-reforms might explain the popularity of *Poor Economics* in philanthropic circles: it legitimizes the work of foundations, while not tackling issues of inequalities, tax evasion (i.e. social budget cuts), and the extractive power of many multinational firms in global value chains, that could make some billionaires uncomfortable. See for instance Kohl-Arenas's (2016: 16–17) descriptions of the ways in which philanthropy cleaved "questions of production, labor, and institutionalized structural inequality from the moral and behavioural explanations of poverty."

micro-structures (*the structure of the program, life structured by goals, the structure of banks*). The two other instances deny the importance of macro-structures:

> What these two examples (the nurses and the school committees) illustrate is that *large-scale waste and policy failure* often happen *not because of any deep structural problem*
>
> it is possible to *improve governance and policy without changing the existing social and political structures*. (p. 270)

We observe a comparable configuration for *macro* (only four instances): pp. 165 and 172, and two instances highlighting J-PAL's seminal experiment on deworming, one of the few experiments to have been applied to large populations:

> We may not have much to say about *macroeconomic policies* or institutional reform, but don't let the apparent modesty of the enterprise fool you: Small changes can have big effects. Intestinal worms might be the last subject you want to bring up on a hot date, but kids in Kenya who were treated for their worms at school for two years, rather than one […], earned *20 percent* more as adults every year […]. The effect might be lower if deworming became universal: The children lucky enough to have been dewormed may have been in part taking the jobs of others. But to scale this number, note that Kenya's highest sustained per capita growth rate in modern memory was about *4.5 percent* in 2006–2008. If we could press a *macroeconomic policy lever* that could make that kind of unprecedented growth happen again, *it would still take four years to raise average incomes by the same 20 percent*. And, as it turns out, *no one has such levers*. (p. 272)

This argument compares an increase in revenues, for a generational fraction of the population, to the growth of the whole country. Is this fallacy of composition more honest than the "anecdotes of fruit sellers turning into fruit magnates"? The authors deny the existence of significant macroeconomic levers. This is an argument of authority, one that begs the question: how can it account, for example, for the economic development of several Asian countries—including China—or the negative multiplying effect of austerity policies (Christiano et al., 2011)? The SAP have had a major impact on the lives of impoverished people, on education, health, access to food or transportation infrastructures. It is impressive to detail the lives the poor without mentioning these subjects. We have seen (point 3.1), the extent of problems that are left behind by RCTs. Thus, the textual analysis contributes to assessing their limited scope.

These rhetorical processes magnify the micro and minimize the macro. The J-PAL's rhetoric is manipulative, not because it attaches importance to behaviors and microeconomic devices: they are certainly fundamental. It is manipulative because it implicitly rests upon a false alternative, between the exclusively all

micro or all macro. This is the *competition of evidence and not in linking evidence.* As Revel (1996: 12) demonstrated, "the problem is not as much about comparing top and bottom, large and small, as it is recognizing that a social reality is not the same according to the level of analysis that one choses." It is therefore the comparison of observation levels that is illuminating: micro, meso and macro are all essential.

8.6 Concluding Remarks: Persuasive but Poor Narratives

"The sources of credibility are threefold [….:] sagacity, virtue, and goodwill."
Aristotle (*Rhetoric*, book II, chapter I, paragraph 5)

8.6.1 An Original Combination of the Three Pillars of Rhetoric

Poor Economics proposes an original combination of the three pillars of rhetoric. First, the *logos* is a discourse that refers to statistical evidence, to an argumentation that reunites all the attributes of scientific rationality (demonstrative argumentation, experimental results, graphs, bibliographical devices). However, this *logos* is inextricably and nimbly linked with *pathos.* Numbers are associated with emotional anecdotes and some numbers, such as 99 cents, are iconic. Concerning the graphs, which at first seem abstract, they are also personified and told through anecdotes and graphic narrations. Furthermore, the authors use multiple figures of speech (metaphors, synecdoches, metonymies, antithesis, oxymorons, etc.) recalling emotions. This is all the more impressive as they explicitly discard the use of pathos and anecdotes and present themselves as the dispassionate voice of science, of hard numbers. Last, the ethos is indeed not forgotten. The authors come across as being endowed with wisdom (*phronesis*), virtue and excellence (*arete*), as well as benevolence (*eunoia*). Here again, they combine emotional qualities (proximity and kindness to the poor), and rational qualities (prestigious scientific background, high standards of rigor). This reinforces the authority of their statements all the while making them, via personal anecdotes, friendly and familiar to the reader.

This textual analysis of *Poor Economics* demonstrates that Banerjee and Duflo have managed to very effectively amalgamate elements that are often considered antagonistic, including by the authors themselves: objectivity and subjectivity, abstraction and personification, numbers and storytelling, rationality and emotion. The textual analysis has equally brought to light two transversal rhetorical schemes: the middle ground between two extremes, and the all-micro rhetoric. They also recall figures of speech as well as manipulative components. This rhetoric has "fascinated and convinced" a Nobel prizewinner, like Robert Solow,

whose macroeconomic posture is, however, the polar opposite of the authors' posture.[14] This canny, often manipulative rhetoric should not overshadow the thinness of its storytelling, the extent of blind spots in RCTs and the danger of having only these "on the menu" (Ravallion, Chapter 1, this volume).

8.6.2 The Capacity to Amalgamate Different Audiences around a Common Content

The rhetorical processes in *Poor Economics* have another important characteristic: they "speak" to very diverse audiences, from a Nobel prizewinner like Solow, to NGOs, to policy-makers and to the *lay public.* The J-PAL does not have multiple discourses that vary across audiences. The core messages are profoundly similar across audiences, from J-PAL online courses to *Policy Briefcases*, from *Poor Economics* to articles in top five journals. Of course, the form of the discourse is modulated for each audience: statistical techniques are central in academic articles and rare in popular science publications. Nonetheless, the argumentative line remains generally the same.

This is a crucial difference with other economic currents. Thus, neo-Keynesians like Krugman and Stiglitz use widely accessible discourses in their popular science works and blogs, discourses that incorporate interdisciplinary and political economic dimensions. This is less true for their academic productions: they are essentially modeled according to extended standard theory (introduction of market imperfections). There is, in them, fundamental differences in the content of the discourse according to the audience. Let's examine the example of Debreu, promoter of an "economic theory in the mathematical mode" (i.e. topological math). After receiving a Nobel prize in 1983, he was hard pressed to talk about the real economy when journalists urged him to do so. But the popular science discourse was, for him, impossible to formulate.[15] Symmetrically, the general public, like policy-makers, could not figure out his publications, which are written in a hermetic mathematical language. *Poor Economics* thus manages to remarkably succeed in overcoming oppositions between scholarly and everyday worlds, oppositions that can be very strong for many economic discourses. This *comparative rhetorical advantage constitutes a decisive factor in the J-PAL's success in attracting students, researchers, journalists, and financing streams*. This is a central element of their expanding business model. This rhetorical element had not been unearthed so far by analyses of the disciplinary success of RCTs,

[14] http://www.pooreconomics.com/about-book/what-others-are-saying.

[15] This episode was reported to me by Alain Desrosières who himself heard it from one of Debreu's daughters.

enlightening other important factors in the political economy of RCTs (Labrousse 2010; Jatteau 2016; Bédecarrats, Guérin, and Roubaud 2019).

8.6.3 Poor Narratives

The authors' make ubiquitous and discretionary use of short stories, despite their explicit discard of anecdotes. This becomes less paradoxical when considering their ethical, social marketing, didactics, and persuasion functions. These brief life stories additionally have a hidden, yet crucial heuristic role in making sense of experimental results and to open the experimental black box. Nevertheless, these anecdotes are thin and ancillary narratives. To fulfill a truly heuristic role, anecdotes would need to be enriched by dense and rigorous narratives, the rules of which were explained in economics by Dumez and Jeunemaître (2005). It would then be a question of, not only marginally but explicitly, combining quantitative and qualitative approaches, in particular ethnographic ones (Morvant-Roux et al. 2014). Experimental designs and interpretations, that are more relevant to the material, social. and cultural environments of the societies in which the experiments are conducted, could thus be developed.

However, such an evolution towards mixed methods has little chance of taking place other than in the margins of economics. Rhetorical studies show the importance of the *audience* in the determination of the form and the content of discourses. Yet, the most fundamental audience of the J-PAL, the one by which they measure their productivity (the number of publishable units per experimentation, Jatteau 2016), the audience that creates careers, is that of the top economic journals. Also these journals almost exclusively value quantification.

Acknowledgement

The author wishes to thank Thierry Guilbert, Stéphane Longuet, Robert Picciotto, Jonathan Morduch, and the editors for their precious suggestions. The usual disclaimer applies.

9
Are the "Randomistas" Evaluators?

Robert Picciotto

9.1 Introduction

The "randomistas" envisage a bright future for development theory and practice through the patient accumulation of experimental evidence at the level of individual interventions. For the MIT Poverty Action Lab's charismatic co-founder (Esther Duflo) and 2019 Nobel Laureate, a new age of scientific progress in the social domain beckons. Thus, she famously announced during a World Bank Conference on the evaluation of development effectiveness: "Creating a culture in which rigorous randomised evaluations are promoted, encouraged and financed has the potential to revolutionise social policy during the 21st century just as randomised trials revolutionised medicine during the 20th" (The Lancet, 2004).

Is this a realistic remit for RCTs—or a manifestation of magical thinking? Since the turn of the century, a huge increase in the use of RCTs has taken place in the development sphere: in a relatively short time, they have achieved dominance in a fashionable social research niche—development impact assessment: the annual publication of experimental and quasi-experimental development evaluations has grown. It is currently plateauing at its 2012 peak of 400–500 studies a year, a remarkable level. Out of 4600 records of published evaluations in June 2018, only 132 experimental and quasi-experimental evaluations were published before 2000 (Cameron et.al. 2016).

About 62 percent of impact evaluations included in the repository of the International Initiative for Impact Evaluation (3iE), used only RCTs and another 5 percent a mix of RCTs and quasi-experimental methods. The balance of about a third relied exclusively on quasi-experimental methods. To be sure, RCTs still account for less than half of the articles in general interest economics journals and less than a third of those in the top-five development economics journals (McKenzie 2016). But two-thirds of the growth in the number of development economics articles published by these journals between 1990 and 2015 was accounted for by RCTs (Banerjee, Duflo, and Kremer 2016). What then explains the rapid spread of RCTs and what does their fulsome embrace by elite universities, philanthropic foundations, and the aid establishment portend for the future of the evaluation enterprise?

Robert Picciotto, *Are the "Randomistas" Evaluators?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0011

In this chapter, I first relate the strong hold that RCTs exert on the public imagination to the deep historical roots of experimentalism. Second, I show that the widespread claim that RCTs constitute a gold standard in evaluation practice flies in the face of the hard-won consensus of the evaluation community, as well as robust evidence that independent evaluation reliant on methodological diversity constitutes good practice. Third, I observe that despite their limitations, RCTs are favored by power-holders who pay the evaluation pipers and call their tune in the contemporary evaluation market. Fourth, I acknowledge that RCTs are making modest contributions to social research. Fifth, I establish that while RCTs are an integral part of the evaluation tool kit, they are not evaluations. Sixth, I conclude.

9.2 Evading the Hard Lessons of Evaluation History

The historical roots of experimentalism are deep. Thales of Miletus, born in the mid-620s BC, first proposed theory-based understanding of natural phenomena in place of supernatural or mythological explanations. Next, systematic approaches to the investigation of nature using deductive reasoning were put forward by Plato and Aristotle. But the institutionalization of scientific inquiry only came into its own in Europe at the beginning of the modern era.

9.2.1 A Faith-based Commitment

As experimentalism became a feature of the scientific method, it evinced a great deal of controversy and only acquired public legitimacy when celebrated as a revival of innocent religion. Appeal to divine sanction was mandatory. It was mobilized to validate the basic tenet of the scientific method according to which positive verification is the only authentic test for knowledge creation and accumulation. Through systematic reconsideration of biblical texts, John Milton and his disciples provided compelling reinterpretations of the Creation. Eventually, their reformist conception of religious faith gave respectability to experimentalism (Picciotto 2011).

A fundamental reconfiguration of the relationship between religion, experimental science, and the public sphere ensued. For Francis Bacon and his Royal Society disciples, careful observation and measurements uncorrupted by dogma were legitimized by the advent of a new strain of Christian apologetics that instructed the public as well as scientists and scholars to secure evidence of divine wisdom through direct scrutiny of the natural order. Eventually, positivism extended the experimental approach to human society by asserting that for the social sciences as for the physical sciences only knowledge that is testable, cumulative, transcultural, and independent of the observer is valid.

Thus, faith in experiments became part of religious doctrine until modernity emerged and disenchantment of the world took hold (Weber 1958) and experimentalism was adopted without reference to any deity. But its sacred features lingered in the public mind. Indeed, they were celebrated by Auguste Comte, the founder of sociology, who developed a "religion of humanity" inspired by positivist principles. Unshakeable public faith in the superiority of the experimental approach has since proved resilient even though its core philosophical assumptions have been discredited.

9.2.2 Weak Philosophical Foundations

By now the epistemological stance favored by RCT advocates has been decisively rejected by social scientists. They no longer endorse the logical positivist tenet according to which invariant generalizations about human relationships can be asserted outside a specific cultural context. Thus, Durkheim initially argued that sociology was tasked with creating its own distinctive approach rather than replicate the methods of the natural sciences.

Max Weber further distanced himself from narrow positivism by suggesting that the complexity of human interactions was such that the social sciences can only uncover causal relationships among hypothetical simplifications of social phenomena. The gap between the social and the natural sciences was gradually widened by critical theorists and historical materialists such as Karl Marx, Theodor Adorno, and Jurgen Habermas. Their competing theories converged on the proposition that the natural and social sciences are ontologically distinct.

Next, Thomas Kuhn articulated the view that theory choice in science depends on paradigmatic considerations that go beyond observation. Post-modern critics went further and sought to debunk the scientific method altogether by promoting the view that all experimentation is subjective if not retrograde, especially when it concerns society. Inevitably such advocacy orientation verging on irrationality left the deconstructionists open to sharp criticism and charges of subjectivity and bias. But by then deep skepticism about evaluative claims that do not make their social purpose explicit had become widespread and positivism especially in its utopian form had lost its luster.

Science is no longer perceived as the ultimate arbiter of social policy and belief in human progress inevitably fueled by technological development no longer holds sway. Probing the interface between power and knowledge, social inquiry geared to communicative action in the public sphere has become a privileged way of using evaluation to promote the public good. But the view that there is a single reality that can be conclusively identified by observation even in the absence of a theory has been discredited. Thus, Karl Popper has shown that in the natural as well as the social world, all scientific research is shaped by the hypotheses held by

investigators and that all theories are mere conjectures subject to refutation: while reality exists, it is only experienced indirectly and imperfectly.

On the other hand, the constructivist belief according to which reality is a pure social construct remains a fringe philosophical stance. A broad-based consensus holds that while experiments are critical to scientific progress, the only valid inference that can be legitimately drawn from them is the refutation of predetermined causation theories. From this perspective, rational decision making in the public sphere can only be guided by plausible albeit fallible, context dependent knowledge derived from rigorous reality testing, scrupulous self-criticism, peer critique, and principled debate.

9.2.3 Resilient Loyalty

Whereas logical positivism has lost its lustre in philosophical circles, it still evokes intense loyalty in parts of academia. Thus, RCT advocates take the view that experimental designs are the *only* scientific basis of ascertaining causation or attribution. Such an extreme position is untenable since biology, geology, astronomy, epidemiology, the forensic sciences, etc. all testify to the proposition that causation can be established without randomized control trials. To quote Lant Pritchett, "if experimentation were the hallmark of science, there would be Nobel prizes for alchemy and not for the physics of astronomy."[1]

Careful observation and measurement can prove or disprove a theory about the natural world without randomization. For example, the prediction of the deflection of light implied by the general theory of relativity was first confirmed by Arthur Stanley Eddington from his observations during the Solar eclipse of May 29, 1919. More recent tests using radio interferometric measurements of quasars passing behind the Sun have more accurately and consistently confirmed the theory.

Similarly, RCTs are not needed in the administration of justice. Investigatory techniques, contestability protocols and rules of evidence are considered sufficiently rigorous to penalize, jail, and in some jurisdictions execute individuals convicted of a crime. Nor are randomized designs flexible enough to embrace the diversity of issues of concern to social researchers, the variability of operating contexts or the complexity of development interventions. Qualitative approaches are essential in pursuit of answers to development dilemmas and challenges.

But the "randomistas" are true believers. They enjoy moral certitude and do not readily accept evidence that contradicts their revealed truth. They exclude other perspectives, prefer to associate with other believers and seek to overcome

[1] Personal communication.

resistance by non-believers through exclusion. A defining characteristic of fundamentalism is that the source of legitimate truth lies in the past: fundamentalists frequently refer to sacred texts and sacred figures. Similarly, radical RCT proponents draw authority from the intellectual contributions of early evaluation pioneers while setting aside the lessons that emerged as the evaluation discipline matured.

According to Alkin (2004), all evaluation doctrines currently on offer can be classified by the extent to which they focus on methods, uses or valuing. He metaphorically positions the major evaluation models that vie for influence under the big tent of the evaluation discipline on three main branches of a bushy evaluation theory tree. Experimentalism occupies a prominent position at the very base of its methodological branch: it was present at the creation of the evaluation discipline.

9.2.4 Evolving Conceptions of Evaluation

Specifically, evaluation pioneers concerned with social programs conceived evaluation as a transmission belt between the social sciences and decision makers.[2] Thus, Donald T. Campbell, the methodologist of the *Experimenting Society,* visualized public interventions as policy experiments. Sharply focused on the elimination of bias in social science inquiry he touted the experiment as "the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition" (Campbell and Stanley 1963).

The "randomistas" still endorse this view even though Campbell eventually reconsidered and nuanced his rigid methodological stance. Given the disappointing results of experimentalist studies in the social policy arena, he revised his negative assessment of qualitative methods and recognized that the identification and elimination of potential claims to causality and the interpretation of side effects of public interventions inevitably require expert qualitative judgment so that in order "to be truly scientific, we must re-establish the qualitative grounding of the quantitative" (Campbell 1974).

Thomas Cook built on Campbell's ideas by focusing on contextual factors and how they affect classical experiments. He developed quasi-experimental techniques designed to overcome difficulties associated with experimental control. He also stressed the importance of consultation with evaluation stakeholders. Similarly, Peter Rossi and Carol Weiss while recognizing the attractiveness of

[2] The advent of the evaluation discipline also coincides with the origins of the development enterprise—a time of optimism when the swords of World War II were turned into ploughshares by the victorious allies.

controlled experiments to eliminate selection bias made seminal contributions to the methodological field by linking the program logic underlying public interventions to theory-driven evaluations.

Lee J. Cronbach's intellectual journey led him further away from a wholesale commitment to randomized field tests. It culminated in a fulsome rejection of classical experimentalism. Ultimately, Cronbach came to the view that only simplistic go/no go decisions are influenced by randomized tests whereas the provision of useful evaluation data for instrumental use requires the exploration of a broad range of relevant issues rather than a narrow focus on the necessarily restricted set of questions amenable to randomized control trials.

Eventually Cronbach's interest in enlightened policy making through evaluation led him to question the external validity of randomized control trials. He ended up doubting whether robust generalizations about human behavior can be secured through social research and he advocated modesty and restraint in the formulation of policy recommendations (Cronbach 1982). Similarly, Robert Stake (2010) who started his evaluation career as a positivist and a mathematician became increasingly disenchanted with the potential of measurement and formal modelling in the assessment of social programs.

9.2.5 Back to the Future?

Actual or feigned ignorance of evaluation history has condemned the development industry to repeat it. The debates that true RCT believers have fomented are not new. Intense conflicts between advocates of quantitative and qualitative methods had long fractured the evaluation and social research world until they were decisively resolved in the 1990s. Following fulsome debate and in the light of numerous publications, almost a decade before MIT set up its Poverty Action Lab. the 'paradigm wars' had been conclusively resolved to the satisfaction of most social researchers and evaluators: both qualitative and quantitative methods have their use, i.e. mixed methods have the edge (Datta 1994).

The back to the future resort to randomized control trials (RCTs) in international development since the turn of the century is therefore paradoxical. The consensus of expert opinion is that methodological diversity adapted to the context trumps rigid adherence to a single evaluation model. RCT advocates have chosen to ignore this hard-won consensus. They remain stuck to the utopian constructionism of the early evaluation pioneers. They allege that field experiments are uniquely equipped to provide rigorous aid effectiveness tests and to generate science-based development knowledge. Never mind that development policymakers have long appreciated the demonstrated capacity of independent evaluation armed with mixed methods to promote self-assessment, track performance, reflect on the lessons of experience and induce reconsideration of misguided

development policy approaches (Grasso, Wasty, and Weaving 2003). RCTs come centre stage in economic research.

Thus, disdain of evaluation history and doctrinal objections to qualitative evaluations underlie the increased popularity of RCTs in development. The failure of macro-economic social research to satisfy aid sceptics facilitated the incursion of micro-economists in the development economics market. Specifically, the ascent of RCTs was linked to disillusion about the capacity of macro-economic methods to generate valid policy prescriptions for the aid industry: a cottage industry of policy research studies grounded in cross-country regressions had generated diverse and contradictory findings regarding the aggregate impact of aid (Tarp 2009).

Macro-economic research could not identify the robust correlations between aid volumes, policy prescriptions, and economic growth that policy-makers were seeking. This is unsurprising: aid works sometimes and fails other times. Context matters, and aid goals vary. Development is not only about growth. The technological and capacity building benefits of aid cannot easily be captured by macro-models. Aid channels, instruments, and modalities matter. So do social and institutional contexts.

Yet, at a time of turmoil in the aid establishment the ambiguous results contributed to public despondency about the usefulness of macro-economic research to ascertain aid impacts. For example, a literature seeking to explain the discrepancy between country-wide aid impacts and project level studies (labelled a "micro–macro paradox") emerged. It threw further academic doubt regarding the account of aid outcomes reported by the development agencies' evaluation functions. Development performance findings, even though based on transparent qualitative methods, were suddenly judged unreliable by experimentalists for whom only quantitative methods are valid tests of attribution.

9.2.7 Micro-economists Enter the Aid Effectiveness Fray

In a tumultuous intellectual environment, two warring factions—the aid optimists led by Columbia University professor Jeffrey Sachs (2005) and the aid pessimists led by William Easterly of New York University (2007)—engaged in intellectual jousts that produced more heat than light, undermined public trust in development assistance and created a strategic opportunity for young economists based at the Massachusetts Institute of Technology (MIT). The necessarily inconclusive outcome of social research findings shifted the focus of the aid effectiveness debate from the abstract plane of macro-economics to the gritty playing field of micro-economics.

Staying clear of grandiose generalizations, the "randomistas" championed a fresh approach focused on a clinical examination of specific development

interventions. Never mind that development evaluators had always been intent on verifying whether the assumptions held by aid practitioners "worked on the ground" at project, sector and country levels. Their work was summarily dismissed by those who postulated that only experimental methods are valid despite ample evidence that qualitative development evaluations have long been and remain essential instruments of due diligence in aid administration.

The indeterminacy of macro policy research results combined with superficial criticisms of qualitative evaluations was fuelled by aid scepticism. Remarkably, the rich evidence of project evaluation studies and the extraordinary success of the development enterprise in many emerging market economies were dismissed as irrelevant on the dubious premise that attribution cannot be established without experiments.

9.2.8 Experimentalism Redux

The combination of alleged scientific rigor, studious ideological neutrality, and "can do" pragmatism proved irresistible. It quickly garnered the enthusiastic support of international philanthropic foundations intent on making their mark on the development scene. With financial help from the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation, an Evaluation Gap Working Group was assembled by the Centre for Global Development (CGD) in 2004. Its underlying rationale was that billions of dollars and thousands of aid programs had been devoted to health, education, and other social sector outcomes without studies that could determine without ambiguity whether they actually "worked."

The Working Group Report (Centre for Global Development 2006) brushed aside the rating system that development evaluators used to assess the effectiveness of aid interventions. It asserted that the results of traditional evaluations lacked validity since they did not address the attribution question in a rigorous fashion. A systematic search for soundly based evidence about the effectiveness of development interventions through "scientific" methods was advocated. Only then would adequate evidence be secured to help close ineffective programs and identify approaches to poverty reduction worthy of replication.[3]

Specifically, the report asserted that ascertaining whether "aid worked" required randomized field experiments or quasi experimental methods that approximate the randomization gold standard.[4] Yet, as noted above, the gold standard claim had been found wanting decades earlier. Evidently the lessons of

[3] In the actual world of practice, this vision was never realized.

[4] In medicine, a gold standard test refers to a diagnostic test or benchmark that is regarded as definitive.

the evaluation paradigm war had not been internalized within the silo of development economics so that the momentum generated by the MIT upstarts proved unstoppable. Gradually funding for development research shifted from macro-economic studies to micro-economic assessments of development interventions.

9.2.9 From Social Research to Evaluation

It did not take long for the struggle for RCT dominance within the social research establishment to spill over onto the evaluation scene and to revive the dormant paradigm conflict. Aid evaluators who had only recently joined the mainstream of the evaluation profession were caught unawares. Unprepared for the onslaught they gave ground. They had not been a party to the methodological debates that took place in the mainstream evaluation community in the late 1970s and early 80s.[5] This is how micro-economists committed to experimental methods invaded a territory that had previously been the preserve of development practitioners turned evaluators. Noisy controversies soon erupted in international evaluation conferences so that the risk of a schism in the development evaluation community loomed.

At one end of the spectrum, seasoned development evaluators schooled in qualitative methods viewed the rigor attributed to experimental methods as illusory. At the other end, evaluators who had long sought closer connections with economics, the queen of the social science disciplines, welcomed the new micro-economists' forays into the evaluation field and advocated close collaboration with them. Following extended deliberations, broad agreement was reached on a methodological guidance document (Leeuw and Vaessen 2009). It acknowledged the frequent superiority of experimental designs to ascertain attribution. But it rejected the hypothesis according to which randomized control trials constitute a gold standard. Instead, it favored mixed methods adapted to the unique needs of specific evaluations.

A Solomonic judgment had once again inaugurated a truce among the contending evaluation parties. But the consensus did not extend far beyond mainstream evaluation circles. In the social research world, and for major evaluation users, misunderstandings still linger, and tensions remain. Evidently, the conflict has been frozen rather than resolved. What then is the consensus of expert opinion regarding RCTs?

[5] The intra-evaluation methodological conflict flared again briefly in late 2003 in the United States when the Department of Education ruled that experimental methods would be privileged in its evaluation funding.

9.3 The Potential and Limitations of Experimental Methods

In the right circumstances and in expert hands, experimental methods provide an estimate of the results that would have been observed had the intervention not taken place. They do so by seeking strict comparability between control and treatment groups through random selection of beneficiaries and non-beneficiaries drawn from the same population through an explicit chance-based process (e.g. a roll of the dice; a roulette wheel or a random number table).

Unbiased allocation means that the probability of ending up in the control group or the treatment group is identical. This RCT feature is intended to address the issue of *selection bias* which arises when a comparison of impacts on two very different sets of beneficiaries ends up falsely attributing the observed results to the intervention even though different known or unknown characteristics of the treatment and non-treatment groups may have been at work.

This includes frequent cases where those who access the program are richer, more powerful, more motivated or more educated. In principle, truly random assignment to the treatment and non-treatment groups from the same population helps ensure that, except for chance fluctuations, the impact of the intervention can be reliably ascertained by comparing outcomes among the two groups by ensuring all the other factors that may affect outcomes are identical except for stochastic errors.

To ascertain the reliability of testing, statistical techniques are available to determine the range of confidence that one may safely attribute to the result (i.e. the role that pure chance associated with the randomization process may have played). Thus, RCTs enjoy the additional advantage of allowing evaluators to establish a measure of statistical significance to evaluation findings.

9.3.1 The Limitations of RCTs

As Martin Ravallion's chapter makes clear (Chapter 1, this volume), the claim that any difference between treatment and comparison groups outcomes can only be due to the intervention is inaccurate. This is because only if the treatment group and the control group and the process that affects each are strictly identical (except in terms of cause and effect) can confident inferences be established. Yet, sampling errors are inevitable and internal validity may be further jeopardized by latent and unobserved causal factors that were ignored when constructing the treatment and control groups.

Such statistical pitfalls are not often acknowledged by RCT advocates and they cannot always be addressed conclusively at reasonable cost. What then is the applicability of randomized control trials for assessing the impact of development interventions? They can be useful if their hazards are recognized and dealt with.

However, they are not always appropriate. They only focus on a single policy parameter whereas most development interventions are driven by complex theories of action and change and aim at many policy goals. They also assume that interventions treatments are fixed and stable while in the real world they are flexible and adaptable.

RCTs are redundant when no other plausible explanation for the results observed is available. They may not constitute a feasible option. For example, it is not possible to randomize the location of infrastructure projects (Ravallion 2009a). Nor are experimental methods feasible when no untreated target group can be identified, e.g. when an intervention is intended to be universal (the imposition of a legal limit for alcohol consumption, a civil service reform program, the liberalization of an import regime, etc.) or when the intervention design is flexible and adaptable to changed circumstances (Lensink 2014).

Nor is external validity the *forte* of experimental methods. Even where experiments are appropriate, they may not meet the needs of policy makers who are vitally concerned not so much with what happened in a trial experimental sample but with whether they are likely to keep working in a diverse, complex and volatile implementation environment (Cartwright and Munro 2010). Program size, structure and context matter a great deal in shaping the outcome of development activities.

The case for observational and qualitative studies also lies in the fact that only experiments within which a plausible theory is embedded are worth carrying out. Thus, systematic reviews that aggregate conditional cash transfer study findings without taking account of differential demand elasticities are close to meaningless. In order to achieve evaluation quality, a deep understanding of how a program operates in its unique context is critical and the theory on which the finding is predicated must be specified. Securing an adequate understanding of causal relationships and identifying the rival explanations that need refutation call for substantive knowledge of the intervention, its design, its implementation protocols, and the incentives of program participants and beneficiaries.

Even where experiments to establish attribution make sense, they require superior skills, large studies, large samples, and specialized quality assurance arrangements. These prerequisites are not always available in the development sphere. As a result, RCTs may not translate into an economic use of scarce evaluation resources. They may also inhibit resort to cheaper and more effective evaluations and hinder fulsome participation of aid recipients in the evaluation process by shifting the control of econometrically sophisticated impact evaluation to well-endowed universities, and think tanks located in developed countries.

9.3.2 Ethical Concerns

RCTs address selection bias where persons who access the programme are richer, more powerful, more motivated or more educated. Random assignment to the

treatment and non-treatment groups from the same population ensures that, except for chance fluctuations, the impact of the intervention can be reliably ascertained by ensuring that all factors that may affect outcomes are identical except for stochastic errors. Randomized control trials also provide evaluators with a measure of statistical significance to evaluation findings.

These are formidable advantages. But experimental methods almost invariably raise ethical concerns that are not often acknowledged by the "randomistas." Depriving members of the control group of a useful treatment based on a selection process perceived as capricious and arbitrary can be discriminatory and may even be illegal. In some jurisdictions, comparison group members are not allowed to receive any treatment that is less than the best currently available.

Nor is it usually considered ethical to induce members of a treatment group to participate in an intervention that may have negative side effects. Paradoxically, informed consent procedures used in such cases may introduce the very selection bias that the method is supposed to avoid so that blind experiments must be used. Even then one cannot eliminate the subtle effects that experiments may induce on the treatment and non-treatment groups (Hawthorne and John Henry biases).

9.3.3 Unintended Effects

Privileging public interventions that are evaluable through experimental methods encourages the selection of simplistic programs and projects that may not be fit for purpose and/or promote avoidance of critical evaluation questions by favouring questions that can be tackled through randomization. RCTs on their own cannot tackle the "why, who, and so what" questions.

Most high-level policies, programs, and projects that are now privileged by international development agencies are not evaluable through randomized treatment. All this means that randomization is mostly suited to narrow questions or simple projects with easily identified participants and non-participants and where spillover effects are not likely to bias the results. It is poorly suited to the evaluation of complicated or complex programs in unstable environments. Yet, this is where knowledge gaps are the deepest.

9.3.4 There Are Alternatives

Many evaluators go through their whole career without ever using a randomized control trial. In part this is because other methods are better equipped to address issues of *why* interventions succeed; *whether* design or implementation problems explain observed intervention failures or *who* among development partners is responsible for observed outcomes. They involve participation, observation, analysis of text-based information, village meetings, open-ended interviews, etc.

Of course, qualitative data collection requires careful coding and systematic quantification to be econometrically analyzed. Qualitative methods guided by theories of change examine what has happened and why. They are better equipped to determine the reasons for success or failure of achieving intended effects (and the extent and nature of unintended effects). They help to discriminate between design issues and implementation problems.

Whereas experimental methods are shaped by data, qualitative, theory-based approaches are shaped by the questions of interest to stakeholders and the assumptions embedded in program and project interventions (Bamberger, Rao, and Woolcock 2010). Finally, a wide variety of tools exist to simulate a counterfactual short of randomization. The listing that follows is only indicative of the wealth of methods and tools available to evaluators. It is not meant as an assessment of their respective strengths and weaknesses in diverse evaluation contexts.

Regression and factor analysis: Regression analysis is used to ascertain the extent to which various characteristics of the context and the beneficiaries of an intervention explain the variations in outcome effects. The balance is attributable to the program on the assumption that all rival explanations have been factored into the model. *Regression discontinuity* compares the effects of treatment on subjects selected according to a criterion (e.g. expert rating of subjects on their likelihood of success or their need for the intervention). It compares the effect of the treatment just above an eligibility cut-off point with those just below.

Quasi-experimental designs: Where randomization is not feasible, one may simulate it through *quasi-experimental* designs. The individuals included in treatment and non-treatment groups in the different are *matched* to ensure that they are similar with respect to the characteristics that may influence the outcome. Statistical adjustments are available to help ensure that the two groups closely resemble each other with respect to these relevant dimensions.

Multivariate statistical modelling: Designed to take account of all postulated relationships among treatment and non-treatment variables the model should be capable of explaining the differences between the two groups at the initial stage so that the differences observed at the post-treatment stage can be netted out statistically. But this approach has problems of its own: it assumes not only that the model has captured accurately the relationships among variables but also that all factors that explain the pre-treatment differences have been identified.

Participatory approaches: Qualitative impact assessment relies on the voiced perceptions of actual or potential beneficiaries, expert observers and/or decision-makers. Color voting facilitates principled debate by displaying stakeholders' opinions through colored presentations of their votes (or scores) on clearly formulated questions about the intervention. Concept mapping involves the use of flip charts and cards (or data processing software) to obtain a graphic image of stakeholders' perceptions of the potential impacts of a development intervention.

It uses skilled moderators to engage a representative group of stakeholders who are knowledgeable and committed to participate.

Surveys and sampling: Survey data collection and interpretation, structured or semi-structured interviews, focus groups, and other methods of involving beneficiaries can illuminate what works, doesn't work and why. Where large groups of citizens or beneficiaries are surveyed data collection and interpretation calls for effective sampling strategies.

General elimination methodology: Michael Scriven (2008) has proposed an alternative to RCTs inspired by criminal investigation techniques that focus on motives, means, and opportunity. This general elimination methodology requires a survey of the literature and/or consultation with individuals who possess tacit expertise relevant to the intervention domain. The process starts with a systematic listing of possible causes that pertain to the intervention. Next, a list of the modus operandi for each possible cause is constructed. This is followed by a detailed examination of the facts of the case. Only the causes "left standing" are retained as potential explanations.

Expert panels: Using expert panels of independent specialists familiar with the domain of the intervention can be useful in conjunction with other methods especially where the evaluation team does not include subject matter specialists or senior evaluators. Panels can be used to assess whether observed impacts are in line with what may be reasonably expected in a specific context. The validity and reliability of expert panels' judgments can be enhanced through a *Delphi process* that consists in consultation procedures with the individual experts without any prior consultation among them.

Benchmarking: Benchmarking uses key performance tests to judge impact through comparisons with good or best practice observed in similar circumstances. Internal benchmarking identifies and seeks to replicate good practices observed within a program. External benchmarking compares the impact of an intervention with that of a similarly situated initiative perceived to have achieved standards of excellence.

9.4 The Current Evaluation Market Favors RCTs

Given that the overwhelming consensus of the evaluation community has come to recognize the severe limitations of RCTs, what explains the remarkable ascent of experimentalism in international development evaluations? Evidently, irrespective of expert opinion, evaluation policy practice tends to reflect dominant interests in society. Accordingly, the evaluation concepts most influential at any one time reflect the mental models that drive power-holders' decisions.

The resulting dynamics have been aptly captured by Evert Vedung (2010) in his famous model of evaluation diffusion. It portrays the history of evaluation as a

succession of waves driven by the changing winds of political ideology. Each wave is propelled by the doctrinal tides currently in vogue. It eventually loses energy and once it subsides it leaves behind layers of intellectual sediment that enrich the discipline and shape its contours.

9.4.1 The Waves of Evaluation Diffusion

Experimentalism is emblematic of the first wave and, as noted above the positivist assumptions underlying it gradually lost support and under democratic governments in the United States, a *dialogue-oriented*, constructivist, participatory, and pluralistic wave surged in the late 1960s when the values underlying the domestic war on poverty and international aid coincided. In turn, the political winds shifted sharply to the right in the eighties. As a result, a powerful *neo-liberal* third wave swelled and engulfed the evaluation discipline. Imbued with new public management thinking it supplanted the constructivist, dialogical, participatory, and democratic evaluative approaches of the second wave.

We are now surfing a fourth wave. It is *evidence-based* and it takes neo-liberalism for granted. It is goal achievement oriented and it favours quantitative methods. It legitimizes value free evaluation by clothing it in technocratic apparel. It gives pride of place to the achievement of policy goals set by power holders. It thrives on tracking progress through theory free indicators. In this authorizing environment, a technocratic, positivist, utilization-focused evaluation approach highly reliant on experimental methods is consistent with the requirements of an evaluation market increasingly controlled by vested interests.

Paradoxically, the same intellectual environment that aspired to more rigorous evaluation methods brought forth new threats to the integrity of evaluation processes and the validity of evaluation results. According to Ernest House (2014) "because of structural changes in society itself, we have a new set of potential biases, a family of biases that we have to deal with or should deal with." These structural changes include the growing encroachment of emboldened private interests over public affairs. The travails of medical research evaluation (still touted as exemplary by RCT advocates) are emblematic of the risks currently faced by the development evaluation enterprise.

9.4.2 The Lure of Medical Research

Embarking on a social transformation initiative through a development intervention is not the same as administering a pill. This is not to say that scientific work cannot achieve rigour in medical research or that randomization is not the method

of choice to assess attribution in some circumstances. But the pitfalls of medical research as currently practiced should be recognized before mimicking it in the development evaluation domain.

In practice, peer reviewed medical research studies disseminated by the mass media have advertised different conclusions regarding the health benefits of such treatments as the regular intakes of vitamins, taking an aspirin a day, sleeping more than eight hours a night, drinking red wine at every meal, the cancer risks associated with using cell phones, living near a high-power transmission line, etc. Extravagant and sometimes fraudulent claims have slipped through the peer-review process of scientific journals, e.g. one large randomized control trial found that secret prayers by unknown parties can save the lives of heart surgery patients while another proved that it can harm them (Freedman 2010).

John P. A. Ioannidis (2005a), Director of the Prevention Research Centre at Stanford University, has designed a mathematical model for assessing the probability that a medical research finding is true. His landmark article confirms that the probability of hypotheses depends on much more than the confidence interval threshold set at 5 percent by most journals. Specifically, his simulations show that poor selection of the relationship being tested, inadequate power of statistical designs, medical treatments characterized by small effects, diverse sources of researcher prejudice etc. have had a devastating effect on the validity of most published research findings.

Even modest levels of researcher bias (either fed by ambition or conviction) are conducive to misinterpretation of statistical tests, distorted use of evidence and/or misleading presentation of results. Published medical research findings are often demonstrably false. Even highly acclaimed research findings can be untrustworthy (Ioannidis 2005b). Erosion in medical research credibility is due to the capture of medical research by vested interests, a risk increasingly faced within the evaluation world.

Until the 1980s drug research was largely independent of the pharmaceutical companies. This is no longer the case: clinical trials are now controlled by private multinational companies and RCTs do not protect the process from many systemic biases (House 2008):

> New drugs are often tested against placebos (the selected counterfactual) rather than drugs currently in use so that minor variations in existing drugs are often recommended for use even if they are not superior to existing drugs.
>
> Comparisons among competing drugs are not always based on equivalent dosages.
>
> Younger subjects who suffer less from side effects are used for tests even though the drugs are more often targeted to older patients.

Time scales are frequently manipulated, i.e. testing is often of short duration even for drugs taken over a lifetime.

Companies not researchers control data analysis and publication so that findings from negative or inconclusive trials are usually suppressed and reports are written to show products in a favorable light.

9.4.3 Distorted Incentives

In today's evaluation market, power-holders hold the purse strings. Evaluations are not designed and implemented without the fulsome involvement of managers. Distorted incentives and threats to evaluation integrity and independence result from such constraints. Unsurprisingly, RCT studies are favored by vested interests since they steer clear of examining the impact on aid outcomes of inadequate program selection and shoddy management performance.

The medical research record demonstrates that RCTs are vulnerable to misleading selection of comparators, cherry picking of data, biases in reporting of findings, financial leverage, etc. when it is captured by vested interests. Even if the research is carried out by universities most trials are now funded by drug companies under contracts that restrict academic freedom by giving private sponsors tight control over evaluation designs, data analysis, research interpretation, dissemination of findings, etc.

Hence, the insidious capture of medical research by vested interests demonstrates that threats to evaluation validity may originate in lack of independence more than in methodological sloppiness. The bottom line is that medical research practice is not a standard of excellence.

Given that commercial and geopolitical interests are increasingly influential in the international aid sphere, the sobering record of medical research evokes looming risks for development evaluation. Only ethical principles and agreed standards of professional practice stand in the way of evaluation capture by partisan interests.

9.5 Modest Contributions to Development Knowledge

Beyond ascertaining whether individual development interventions "work" as intended, the "randomistas" aim to generate important social research and policy findings. According to the MIT Poverty Action Lab (J-Pal) website: "Randomized evaluations can generate important insights about human behavior and institutions in addition to measuring the impacts of specific programs and policies. The knowledge generated across multiple randomized evaluations on the same topic can help inform decision-making in governments, NGOs, firms, and funders

working to address similar challenges" (Dhalival and Olken 2018). The record suggests that these claims have limited validity.

9.5.1 A Narrow Scope

The theory-free advantage that RCTs enjoy for addressing attribution questions at the intervention level turns into a disadvantage in social research unless they are paired with other methods and build on prior knowledge (Vivalt, Chapter 11). This is because individual RCT studies on their own cannot claim replicability across operating contexts. The statistical hazards associated with sampling severely hinder the transportability of findings outside the context in which the experiments were designed and carried out. This is not only because RCTs do not always give reliable estimates of average treatment effects but also because guaranteeing causality at the intervention level does little to establish external validity of RCT findings (Deaton and Cartwright 2018).

Furthermore, RCTs are methodologically parsimonious and have a limited reach. Since they are mostly concerned with eliminating the selection bias of development interventions, they only address narrow questions about the efficacy of delivery mechanisms for private goods. Public goods, i.e. goods that are non-rivalrous and non-excludable, cannot be readily subjected to randomization.

This means that RCTs are not equipped to tackle critically important development policy issues, e.g. climate change, biodiversity, public safety, intellectual property, etc. For such goods that are at the core of sustainable development policy it is not practical to design experiments that distinguish between those that did or did not benefit from "treatment."

9.5.2 A Paternalistic Stance

RCTs privilege examinations of how aid beneficiaries (i.e. the poor) think and behave. This stance is consistent with the view that poverty is a personal choice rather than the consequence of existing social arrangements and political structures. The "randomistas" do carry out field work in order to construct statistically plausible surveys. But they privilege their pre-existing mental models, and they concentrate on policy tweaks rather than alternative policy choices.

As a result, boosted by the findings of fashionable behavioral economics, they are prone to question the rationality of poor people's choices and rather than examining the social dysfunctions that limit their options and undermine their economic prospects, they focus on how policy-makers can nudge them towards

pre-determined behavioral changes even though such changes may not reflect their individual preferences or their circumstances.

9.5.3 Limited Contributions to Knowledge

Evaluation delivers significant results when it addresses important and pertinent operational questions. Judicious selection of evaluation topics is essential to make evaluation pay. From a utilization perspective, independent evaluation conceived as an organizational learning tool and focused on strategically pertinent issues has major advantages over scattered experimental evaluations carried out in widely different contexts, driven by diverse, often self-interested clients especially when they are implemented by outsiders with limited development experience who are hampered by massive information asymmetries and incentivized by the academic urge to publish.

To be sure, RCTs have contributed to development knowledge when they have addressed a pertinent policy question, when they have drawn on the accumulated findings of the literature and when they have been complemented by observational studies and qualitative methods. Thus, the Swedish academy of sciences, enthused by the on-ground experimentation savvy displayed by the MIT and Harvard economists awarded them the Sveriges Riksbank 2019 prize.

For example, RCTs have helped to falsify the exaggerated claims of micro-credit fervent advocates who had grounded their expectations in case studies that described micro-credit schemes as the key to women empowerment and large-scale poverty reduction. Carefully constructed RCTs in diverse contexts combined with field observations have shown that micro-credit is a useful financial product but that it is not the key to radical social change.

In some instances, micro-loans made no appreciable difference in women's influence on household decisions and spending patterns. Equally the rigid terms and the group lending rules designed to protect the financial sustainability of micro-credit institutions were shown to be poorly adapted to the needs of budding entrepreneurs. Nor did the business training programs tried out by micro-lenders to help borrowers grow their enterprises have a significant impact on their profit or sales (Banerjee and Duflo 2011). In this way, RCTs have helped to debunk some of the fashionable, yet flawed models that have periodically swept over the development scene.

RCT studies have also "rediscovered" well-established good practices in development, including the effectiveness of remedial tutoring in schools and of preventive health care highlighted by the Swedish Academy of Sciences. Similarly, 58 Poverty Action Lab RCT studies provided field evidence backing what experienced education policy practitioners had already concluded regarding the drivers

of increased student enrolment and participation, i.e. reducing (or eliminating) school fees, cutting down on travel times to school, attending to children health problems and providing information to parents about the benefits of education.

Along similar lines, a field experiment in 100 Indian villages validated the findings of prior agricultural extension studies: farmer field days are useful and cost effective in the dissemination of new high yielding varieties. Furthermore, and unsurprisingly, an elaborate experimental study carried out in Kenya confirmed that profit maximization at farm level rather than yield maximization should guide advice about fertilizer applications. It is as if the "randomistas" were looking for evidence that economics has merit or that their favored evaluation instrument "works."

9.6 RCTs Are One Tool among Many

Given these observations, are RCTs consistent with the core principles, purposes, and practices of the evaluation discipline? While definitions of evaluation and evaluation models are legion, most acknowledge the critical role of *value* in evaluation—viz. the concise definition offered by Michael Scriven (1991) that has gained broad based acceptance in the evaluation community: "the process of determining the merit, worth and value of things – or the result of that process". The three dimensions of interest in this definition are interrelated but it is the value feature that most distinguishes evaluation from other types of inquiry.

First, *merit* ascertains performance relative to quality standards. It has to do *with doing things right* to achieve intervention goals, i.e. *efficacy* which is defined in the Glossary of the Development Assistance Committee (2010) as "the extent to which the development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance."

Next, *worth* has to do with *doing the right things*. It refers to the net benefits can be legitimately be ascribed to the intervention taking account of merit considerations grounded in the perspectives of those who are expected to benefit from the intervention and other stakeholders, persons, or entities affected by the intervention. It is about *relevance* defined in the Development Assistance Committee (DAC) Glossary as "the extent to which the objectives of a development intervention are consistent with beneficiaries' requirements, country needs, global priorities and partners' and donors' policies."

Finally, *value* evokes the public interest and it also brings in considerations of economy in the resources used to achieve the intended results, i.e. doing things efficiently relative to other ways of designing and implementing the intervention. Specifically, *efficiency* is defined by the DAC Glossary as "a measure of how economically resources/inputs (funds, expertise, time, etc.) are converted to results."

9.6.1 How Evaluative Are RCTs?

RCTs are an integral part of the evaluator's tool kit and there is little doubt that ascertaining causality of observed results (the fundamental purpose of RCTs) is an integral part of assessing its merit. On the other hand, this restricted approach to evaluation does little to establish whether an intervention is relevant, efficient, or sustainable. Finding out whether an intervention works is not the same as examining whether it was the right intervention, figuring out why it performed the way it did or whether its goals were worth pursuing in the first place.

Program goals, size, structure, and context matter a great deal in shaping the outcome of policy and programs. Even where experiments are the right way to approach attribution analysis the results may not meet all the felt needs of policymakers concerned not so much with what happened in a trial experiment but with whether the experiment is likely to keep working in other contexts or in the future given the wide prevalence of complex and volatile implementation environments (Cartwright and Munro 2010).

Finally, without a theory subject to falsification no advance in knowledge is possible. A deep understanding of how a program operates is needed for high-quality evaluation where the validity of the theory on which the program is predicated must be established. Securing an adequate understanding of causal relationships and identifying the rival explanations that need refutation call for substantive knowledge of the intervention, its design, its implementation protocols and the incentives of program participants and their beneficiaries. Open-ended questions and qualitative approaches are better suited to deal with such issues.

This explains why independent evaluation, grounded in field work, embedded in the organization and carried out by experienced practitioners has proved far more effective than RCTs in reorienting operational processes and in shutting down (Gautam 2000) or restructuring of ineffective lines of development lending (Tendler 1993). Nor is the caricature of internal evaluation as inevitably subservient to institutional self-interest valid, especially where the evaluation function reports to the supreme authority of the organization rather than to operational management and where it is mandated to attest to the quality of self-evaluative processes (Picciotto 2013).

Politicians and civil servants make collective choices about how public resources are allocated and used. They are mandated to secure high value for the bundle of assets assigned to their care. They need to demonstrate that they are doing so responsibly and effectively. Hence the key to the legitimacy of power and authority is a valid and authoritative narrative regarding the creation of public value.

As a *summative* endeavor evaluation examines the results of policies and programs and focuses on the extent to which the authorities who were in charge acted responsibly. The main restoration mechanism to poor government performance is

citizens' voice. Evaluation amplifies it by providing relevant knowledge about public sector performance to voters.

Measuring public value through simple output measures and budget coefficients rather than outcomes and impacts has dominated public sector management. Such indicators leave a lot to be desired. They do not measure results and they can easily be manipulated. Hence, the information provided by public sector managers about their work needs robust validation: independent evaluation in the public sector is what auditing of accounts is in the private sector.

This is where independent evaluation comes in: it is tasked to ascertain reliably whether errors in decision-making were due to circumstances over which decision makers had no control or whether the risks incurred could have been managed better. Fair and objective evaluation contributes to accountability: it ensures that the promises of politicians' and decision-makers in the public, private and voluntary sectors are systematically compared with what is delivered through fair and objective evaluative processes. Relating results to the promises made when a policy or program was launched is part and parcel of the democratic process.

Thus, goal-oriented methods have a privileged place in the evaluator's arsenal. But in this respect, experimental evaluation makes no claim to distinguishing between the effects attributable to the diverse actors invariably involved in policy and program interventions. Yet, most policies and social programs rely on *partnerships* between various government, private sector, and civil society entities to achieve outcomes and impacts. Without assessments of their distinct accountabilities and their compliance with reciprocal obligations, the responsibilities of partners are blurred.

For example, responsibility for failure may be shirked altogether if it summarily attributed to poor partner performance. Conversely responsibility for success may be unfairly captured by a single partner (e.g. a government agency)—whether its contribution to the shared objectives justifies it or not. Lack of adequate evaluation can therefore have deleterious effects on incentives through flawed signals.

Hence, when program or project failure (when it occurs) is ascribed entirely to the implementing agency (irrespective of exogenous influences and of partners' contributions) it induces risk aversion and it may even encourage suspension of programs that fail to meet ambitious goals thus forsaking the opportunity of adapting them so that they can succeed.

It follows that good evaluations take explicit account of partners' distinct accountabilities and reciprocal obligations. Unless performance of various actors is assessed separately to explain outcomes and impacts moral hazard is bound to prevail. Hence high-quality summative evaluation goes beyond answering the question of whether a policy or program works or not, the narrow focus of experimental impact evaluation.

In sum, for an approach often misleadingly touted as strongly supportive of accountability the new conception of impact evaluation using RCTs evades awkward questions about who might be called to account for observed shortfalls between policy and program goals and actual results. By limiting its focus to the *attribution* of effects to the intervention, RCTs fail to address the *contribution* question—how well did each of the individual development partners perform towards the achievement of program or project objectives and what might be done to improve their performance?

9.6.2 Wielding the Right Tools

RCTs are only one evaluation tool among many. As such they should not be allowed to dominate what is first and foremost a creative, analytical, and participatory process. Experimental methods have many statistical features that other evaluation designs cannot easily match in some circumstances. But a threat to good evaluation management is overinvestment in a single technique. A tool can only fulfil the function or functions that it was designed for.

Using the right tools and using them with care and skill is an important ingredient of evaluation quality. Inappropriate methods can sink an evaluation. But threats to the rigor of an evaluation may also result from other factors: sloppy data collection, politically naive evaluations; lack of independence; inadequate evaluators' competencies; failure to focus on utilization; ignoring the context; limited involvement of stakeholders; concentration on unimportant or irrelevant issues; etc.

Well-selected evaluation tools used according to their specifications contribute to the validity of evaluations. They make evaluations easier to compare and facilitate their costing and their planning. They make evaluation findings more credible and predictable. Understanding and measuring the limits of the tools used in context is critical to quality. The inability to connect the detailed design of the evaluation to the priority questions identified at the planning stage explains why many evaluations go bad.

Consequently, understanding of the respective strengths, weaknesses. and limitations of evaluation methods and tools is a critical competency for evaluators. While experimental and quasi experimental methods can in some circumstances illuminate attribution of observed outcomes, theory-based observational studies, and process evaluations that use judicious triangulation of methods are better equipped to answer how and why the observed effects have materialized. It is therefore fortunate that all national and regional evaluation guidelines and standards give adequate weight and credence to qualitative approaches. They stress methodological appropriateness and pluralism rather than doctrinal orthodoxy.

9.7 Conclusions

Experimentalism has deep historical roots. Successfully marketed by academic entrepreneurs, RCTs evince intense loyalty among their practitioners. They promise certainty and rigor in a development enterprise characterized by extraordinary volatility and complexity. Yet, they are hindered by a host of limitations, they are expensive, and they face a host of statistical and ethical challenges. Their underlying epistemological foundations are unsound, their gold standard credentials are invalid and frequent claims that the randomized trial procedures that made their mark in the health sector hold the key to evaluation rigor in the social sector are groundless.

At the level of individual interventions, RCTs only allow attribution conclusions to be drawn for simple interventions implemented in stable environments and they only contribute to generalized policy research when they are part of a cumulative knowledge generation process that also relies on observational and qualitative studies. Other evaluation methods that may or may not be combined with the experiment are available to deal convincingly with the complex questions of the evolving development enterprise.

As evaluations, RCTs only tackle one of the core evaluative criteria (efficacy) that policy and program interventions must meet to be considered effective. They fail to deal with issues of relevance, efficiency, and sustainability that are often more important. Nor do RCTs provide estimates of the distinct contributions of partners responsible for the success or failure of policy and program interventions, a major weakness since accountability to citizens is part of the evaluation remit.

Thus, the "randomistas" are not evaluators since RCTs are not evaluations. But RCTs will continue to play a major role on the development scene since they have become firmly embedded in the academic world, make modest contributions to development knowledge, do not challenge the prerogatives of power-holders, and within their limited scope, meet an effective demand for publicly plausible evidence as to whether development interventions "work." The 2019 Nobel award will further solidify the privileged role of experimental studies in development economics.

Acknowledgement

Lant Pritchett offered judicious comments on a prior version of this chapter but he is not responsible for its errors and omissions.

10
Ethics of RCTs

Should Economists Care about Equipoise?

Michel Abramowicz and Ariane Szafarz

10.1 Introduction

Is lack of resources a good reason for providing the treated and control groups in randomized controlled trials (RCTs) with unequally promising options? Apparently, physicians answer "no" to this question but economists tend to say "yes." Equipoise is an inescapable building block of medical RCTs. Strangely, many economists performing RCTs never heard about it. This chapter fills the gap and investigates how the equipoise principle is formalized in the medical literature, and subsequently whether and how it should be taken seriously into consideration by economists.

Equipoise is defined by Freedman (1987: 141) as a "state of genuine uncertainty on the part of the clinical investigator regarding the comparative therapeutic merits of each arm in a trial." The author considers this principle as "an ethical necessary condition in all cases of clinical research." Equipoise requires that before the trial starts, there is equal ignorance about the benefits and drawbacks of the treatment options. This requirement is grounded in the ethical motivation that any ex-ante preference for a given treatment option would undermine the interests of those who are offered another. And since the typical medical procedure in RCTs relies on double-blind treatment allocation, failing to fulfill equipoise would potentially hurt all the trial participants. In that sense, the equipoise requirement reinforces the 1964 World Medical Association's Declaration of Helsinki[1] stating, among other things, that control groups must receive the best existing treatment. This requirement is absolute, i.e., it applies regardless of the study's specific conditions, including its location.

Yet, the topic of equipoise is still controversial as its practical implementation raises key issues such as the balance between the opinions and preferences of the clinical community, the individual investigator, and the treated patient (Lilford

[1] Carlson et al. (2004) discuss the later revisions of this Declaration. See also the 1982 International Ethical Guidelines of the Council of International Organizations of Medical Sciences for Biomedical Research Involving Human Subjects (CIOMS 2002).

Michel Abramowicz and Ariane Szafarz, *Ethics of RCTs: Should Economists Care about Equipoise?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0012

and Jackson 1995). Evidently, the appreciation of therapeutic merits may vary according to the sensitivity of the people involved in the implementation of the trial, and so lead to ethical dilemmas (Schafer 1982), such as the thorny question of balancing a doctor's duties to her patient and to the advancement of science (Botros 1990). Miller and Joffe (2011) however contest that the stakes are confined within the doctor–patient relationship. The authors place the debate in a wider context that relates to health policy. They weigh the interest of individual patients against the knowledge needed for drug approval. By doing so, they link equipoise to the typical public-health trade-off that opposes individual liberties to social justice (Kass 2001; Childress et al. 2002). In at least this respect, there is a clear connection between the ethics of RCTs in the fields of medicine and economics.

While the medical literature fiercely debates the relevance of various specifications of the equipoise principle, research in economics is still silent on the topic. Of course, RCTs in economics are usually reviewed by ethical committees. Yet, these committees are typically local. Large-scale ethical requirements, including reference to (any sort of) equipoise, are still missing. We aim to break the apparent indifference of economists to an ethical concern that is key for medical experimentation. In line with Baele (2013) and Petticrew et al. (2013) who advocate the development of "social equipoise," this chapter intends to initiate an equipoise conversation within the economic RCT community.

10.2 What Is Equipoise?

The use of human beings as experimental subjects has created difficult ethical problems. The practice of assessing medical treatments with controlled experiments dates back from ancient times, but the equipoise principle is far more recent. It was formalized in the twentieth century following the designing of randomization with placebo control groups and concealed assignments (Di Tillio et al. 2017).

Most modern codes of medical ethics are guided by the principles of the classical Hippocratic Oath devoted to the obligations of physicians to their patients. (Orr et al. 1997; Miles 2005). A central tenet of the Oath concerns the duty of providing the best available treatment. Specifically, if the doctor has good reasons[2] to believe that treatment A is better than treatment B, then she cannot prescribe B instead of A to any of her patients (Shaw and Chalmers 1970). Likewise, she should refrain in participating in any scientific study that would lead to giving treatment B rather than treatment A. This strong restriction imposed by medical ethics can hinder the development of large-scale medical studies based on the comparison of

[2] The doctor's beliefs mix scientific knowledge with subjective experience and personal thinking. The subjective component inevitably adds complexity to formalizing the equipoise principle.

treatments A and B. To address this issue, Freedman (1987) introduced the concept of *clinical equipoise* that embeds the need of sufficient statistical evidence to conclude that treatment A does not dominate treatment B.[3] The underlying idea is to place the so-called genuine uncertainty about the comparative therapeutic merits of the two treatments in the hands of the expert medical community rather than in the hands of an individual investigator (Freedman 1987).

Without affecting the fundaments of RCTs, clinical equipoise helped relax practical constraints that had sometimes dictated stopping studies prematurely when early results indicated that one treatment is better than the other, at least in the short term. Freedman's idea was to leave enough time to the scientific community to build strong evidence based on large studies. Meanwhile, Freedman argues, clinical equipoise addresses, at least partially, the low take-up associated with the reluctance of physicians to enroll their patients in studies they do not feel comfortable about (Taylor et al. 1984).

Overall, the operational principle of clinical equipoise proved itself fruitful to the development of large-scale medical studies by contributing to the design and applicability of RCTs. By so doing, equipoise helps reconcile the rights of study participants and the quest for scientific breakthroughs (London 2017).

The initial steps of the implementation of the equipoise concern into medical RCTs can inform the economists about the challenges and stakes involved. The cardiological clinical and investigational community was the first to implement large-scale medical RCTs aiming to inform the practical dispensing of patient care. One such pioneering study was the Beta-Blocker Heart Attack Trial (BHAT 1982). Beta-blockers are heart drugs developed in the 1960s, whose discovery was rewarded in 1988 by the Nobel Prize for Medicine granted to Sir James Black. These drugs were initially popular for the treatment of hypertension. The BHAT trial randomized survivors of a recent myocardial infarction to either Propranolol, the first widely available beta-blocker, or placebo (Yusuf et al. 1985). The results showed a very significant reduction (7.2 percent vs. 9.8 percent) in medium-term total mortality (the average follow-up was 24 months). To this day, 36 years after the publication of this seminal study, beta-blockers are still the cornerstone of secondary prevention after myocardial infarction. Beta-blockers are also the most active drugs against angina pectoris—the chronic painful condition caused by partially blocked cardiac arteries, a total blockage typically resulting in an infarction—and they significantly reduce mortality in heart failure (McMurray 2010).

[3] Development economists could argue that some poor populations have access neither to treatment A nor to treatment B, so that even treatment B would improve their condition. Section 4 addresses this argument often used to justify the inferior treatments provided to the control groups. We content that comparing life under a controlled experiment to regular life conditions ignores the role of the investigators, who can affect people's behaviors and feelings significantly. In that regard, RCTs in social sciences are on an equal footing with medical RCTs, which should prevent them from adhering to lower ethical standards.

Before embarking on the BHAT, there was genuine equipoise among the clinical and scientific medical community about the potential protection offered by beta-blockade after a myocardial infarction (Nies, Evans, and Shand 1973; Shand 1975). Animal experiments had shown enhanced survival, and patients with angina pectoris fared well with the drug, but there was concern that the associated blood pressure lowering would result in a net harm (Theroux et al. 1974).

Yet, the level of uncertainty about treatment superiority can change during the course of an RCT for several reasons, including the partial results of the trial itself and the publication of meaningful results by other teams. The data and safety monitoring boards (DSMBs) have the responsibility of determining whether "equipoise has been sufficiently disturbed during the course of a trial to warrant stoppage" (Dickert and Emanuel 2015: 31). This is a difficult decision to make since, on the one hand, early stoppage can harm the overall validity of the study, and on the other hand, the continuation of the study with a compromised equipoise can harm its subjects. For example, there were rumors that the Second International Study of Infarction Survival (ISIS-2 1988) ignored the ongoing findings published in 1987 by an Italian competing research team, the Gruppo Italiano per la Studio della Streptochinasi nell'Infarcto Miocardico (GISSI). Precisely, the ISIS-2 trial investigated 17,187 patients admitted in the Coronary Care Unit (CCU) with a working diagnosis of acute myocardial infarction (AMI). The ISIS-2 patients were randomized between Streptokinase, a clot-dissolving drug that was hoped to better the prognosis by reducing the infarct size, and placebo. In the course of the ISIS-2 study, GISSI partial results strongly favoring Streptokinase began to emerge suggesting that ISIS-2 was unfair toward its placebo patients, i.e. the control group, with respect to their treated counterparts. Yet, this episode unfolded between 1985 and 1988, a period during which the equipoise concept was still little known in the medical research community. Ultimately, the ISIS-2 study went on as planned and its results confirmed the substantial mortality benefits of Streptokinase. Nowadays, 30 years after its youth waywardness, equipoise belongs to the core ethical standards of medical RCTs.

10.3 Equipoise vs. Blindness

Even though there is a large consensus in the medical community that equipoise belongs to the ethical standards of the profession, the practical implementation of this principle raises several practical issues. The most basic problem, and perhaps the most challenging one in the field, is how to prove that a given trial satisfies clinical equipoise. Several methods can be used to bring convincing evidence. They include references to previous studies and testimonies of divergences about treatments within the clinical community. Specific conditions can however compromise the implementation of clinical equipoise. This section comments on

two major issues that may sound familiar to social scientists: the lack of experimental blindness and the decision to enroll a patient with a pre-existing condition.

Double blindness has become the norm of medical RCTs. In economic RCTs, by contrast, it is hardly implemented, for alleged reasons of practical unfeasibility. This section shows that, like in economics, implementing blindness is not always feasible in medical subfields, such as surgery.[4] But in line with the medical literature, we content that even with partial or no blindness, the principle of equipoise should be followed. For instance, the GISSI (1987) study was an unblinded RCT "of which protocol specified three interim analyses, at 3000, 6000, and 9000 recruited patients. Results from these were presented to the ethics committee only; a difference in mortality exceeding three standard deviations or an unacceptably high incidence of adverse reactions to SK would have led the committee to call an early halt to the trial" (GISSI 1987: 398). In fact, the absence of experimental blindness makes equipoise an even more imperative requirement.

Nowadays, medical RCTs are typically conducted in a blinded fashion. In a single-blinded study, the patient is unaware of which treatment she receives. Double-blindness means the subject and the field investigator (who treats the patients, follows them and transmits their end-points results to the trial steering committee) are both unaware (Days and Altman 2000). The purpose of blindness is to limit bias, since knowing which treatment is applied could result in unconscious misinterpretations of the results. Also, the subjective experience of trial patients can be colored not only by their knowledge of which drug they are receiving, but also by their feelings about their doctor's expectations. Patients often think that the experimental medication will perform better than the reference drug or the placebo—even though with equipoise they are mistaken—and will tend to report less symptoms when they know that they are receiving the perceived better option. Podsakoff et al. (2003) summarize the potential sources of common biases observed in behavioral research. For example, the bias of social desirability stems from the patients' desire to respond according to perceived social acceptability rather than to their true feelings. To some extent, the same biases apply to medical providers. Importantly, single blindness can favor—often subconscious—investigators' self-serving biases (Camfield et al. 2014), hence the advantage of double over single blindness. This argument is particularly relevant for soft or subjective tested outcomes, which are common in social sciences. More generally, the hardest the end point, the lesser the interest of blindness. In the extreme case, it is difficult to misinterpret death. But even so, when feasible, double blindness has become the norm in medical RCTs.

[4] Young et al. (2004) suggest assessing the feasibility of equipoise-based RCTs in surgery by organizing opinion polls among surgeons in the field. Such polls would gauge the possibility of testing a surgical treatment against another surgical treatment.

While most medical RCTs deal with drugs, some evaluate lifestyle interventions or surgical procedures. For these studies, finding a placebo that is consistent with using double-blind (or even simple-blind) trials is far from obvious. Surgery-based studies raise the tantalizing question of whether to perform a sham operation. Here is the dilemma. First, if the patients in the control group receive no surgery, they obviously know that they are untreated, and the placebo effect is lost. Alternatively, all the patients can be brought to the operating room, resulting in the unoperated patients entering the recovery ward still half-anesthetized and with a fresh surgical scar. The second scenario would undeniably add validity to the trial in that the only difference between the control and treated groups will be the actual surgical treatment, but at the same time the trial will have done real harm to the controls, in having subjected them to totally unnecessary surgical (albeit not therapeutic) and anesthetic risks. One can see the dilemma as a trade-off between individual interest and the greater good, or as equipoise vs. single blind. In social science, where experimental blindness is practically hard to reach, if not impossible, the trade-off should, somewhat paradoxically, play in favor of equipoise.

In addition to the pure ethical rationale, the so-called Hawthorne effect provides an argument showing that the absence of blindness in experimental design reinforces the need for using some equipoise principle. The Hawthorne effect, uncovered by sociologists (Roethlisberger and Dickson 1939), refers to the situation observed in Chicago's Hawthorne Works of the Western Electrical Company from 1924 to 1927. The researchers observed that regardless of whether they tested an increase or a reduction of the plant's artificial illumination intensity, they obtained a positive effect on workers' productivity. The psychological experience of benefitting from a change in working conditions appeared to be more important than the nature of the change itself. Such psychological benefit can only take place in a treated group where people are aware of being granted a change in conditions, and not in a control group where no change is operated. Hence, providing no placebo-type of change to the control group can harm not only the individuals in the control group but also the reliability of the study by leading to falsely positive discoveries driven by the Hawthorne effect. More and more economists address this issue by providing some type of change to the control group.

Another moral tension stems from the decision to enroll a patient with a given condition in a study. The tension opposes again the physician's commitments both to her patient and to scientific progress. According to Weijer, Glass, and Shapiro (2000: 756), "the answer seems to depend greatly on which side of the Atlantic you reside. In the United Kingdom, the individual uncertainty principle is widely endorsed. However, in North America, clinical equipoise—reflecting collective uncertainty—is the dominant ethical basis." The so-called uncertainty principle works like equipoise applied specifically to each patient considering her

individual pre-existing condition. Under that principle, the interests of each patient must be examined individually by the doctor before any enrollment in the study can be contemplated. A potential consequence is poor recruitment in studies, where ultimately clinical equipoise is put at risk. By contrast, the U.S. National Bioethics Advisory Commission (NBAS 2001: 77) prescribes that Institutional Review Boards (IRBs) "determine whether the relation between risks and potential benefits is reasonable. To do so, IRBs should determine whether the procedures meet the criteria of research equipoise […] in addition to being justified in terms of the potential knowledge gain for society. Investigators and IRBs should understand that the term *research equipoise* applies to any type of research involving interventions or procedures that offer the prospect of direct benefit to participants […]." Typically, subjects whose conditions have proven therapies benefit from studies designed as "standard treatment plus placebo" versus "standard treatment plus experimental drug."

10.4 Should Economists Care about Equipoise?

While economists feel comfortable with RCTs in which the control group receives nothing, medical scholars consider that a placebo, which is better than nothing, is no ethically admissible option when the most recent state of science has uncovered a better drug. Following the equipoise requirement, this stance holds even if the better drug is expensive and the tested population cannot afford it.

How come that the ethical debate surrounding the application of equipoise has remained so confidential in the economic RCT community? Due to the relatively recent emergence of RCTs in economics, it could well be that economic research is still in its pre-Freedman stage. This possibility however has little credibility given that economists have largely borrowed from medical experimental designs in drafting their own studies. Researchers from both the medical and economic communities are therefore likely aware of the ongoing debate about unethical medical studies (Halpern, Karlawish, and Berlin 2002), including those organized in developing countries (Gulhati 2004; Jintarkanon et al. 2005; Milford, Wassenaar, and Slack 2006).[5] And yet, economists and other social scientists tend to disregard the equipoise requirement by typically disadvantaging the control group. Take, for instance, the recent debate that opposed Professor Megan Stevenson from George Mason University, who investigates the impact of bail money through RCTs, to the Massachusetts Bail Fund (MBF), which pays up to $500 bail for low-income

[5] For instance, the Indian Council of Medical Research showed concern that trials ensure compliance with ethical guidelines (Chatterjee, 2008). Mudur (2005, p. 1044) quotes Prof. Falguni Sen from Fordham University, New York, saying that "Given the vulnerability of uneducated and poor patients, India has a long way to go in ensuring adequate protection to human subjects."

people. Each side developed a number of insightful arguments, and the following extracts from their Twitter conversation epitomize the essence of the equipoise controversy.[6]

> "RCTs randomly assign a "new treatment" (in this case assistance from the bail fund) to one group while another group receives the "standard treatment" (in this case no help with bail, resulting in pre-trial incarceration). Terrified yet? WE KNOW THE IMPACT OF MONEY BAIL. The research EXISTS: people have better case and life outcomes when they can fight their case from freedom. We believe people when they tell us the difference it made to have their bail posted. It is WRONG to randomly pick some people to receive a lifesaving treatment while sending some people to jail. It is RACIST to engage in this kind of "research" when you know racial disparities in our courts and jails are long-standing and stark." (MBF, March 8, 2019)
>
> "Until you have already served every single client (for the bail fund, (this would mean having the capacity to bail out every single person that needs it) there are some people who are "randomly" not receiving services. An RCT just makes this random process more explicit. For instance, say you only have resources to staff the courts 5 days a week. Randomize which days are staffed. Or say you don't have time to meet with every defendant during a shift. Start by meeting defendants with odd docket #'s, then move on to even #'s if there is time. Both of these methods are RCTs and would create very valuable research! Because although you think you know that your program is extremely effective, you really don't. Are you reaching the clients that need you the most? Those who would suffer were it not for your help?" (Prof. Stevenson, March 14, 2019)

Redrafted along the equipoise line of thought, the argument of MBF is that assigning randomly a bail to the treatment group and no-bail to the control group violates the equipoise principle because the control group is disadvantaged. Prof. Stevenson dismisses this argument on the grounds that previous on-field knowledge is insufficient to assess confidently the superiority of the treatment ("although you think you know […] you really don't"). She puts the search for rigorous scientific validation above the concern that equipoise is not fulfilled.[7] She also refers to the typical rationale of development economists that real life is already unfair/uncertain for RCT subjects ("Until you have already served every single client"). Basically, the idea is that the RCT is important for science and for future generations.

[6] We are grateful to Tim Ogden for having called our attention to this Twitter conversation.

[7] Garchitorena et al. (2019) mention that "many RCTs are carried out only to confirm results of observational studies."

Thus, a realistic possibility is that economists consider their own studies as benign, meaning that organizers don't need to bother about the potentially negative consequences for participants of the control group who fail to receive any benefits of the tested treatment. For instance, the argument goes, failing to supply a farmer with the loan she could need is no big harm. In economic terms, the situation where 50 percent of the farmers receive the loan and 50 percent do not is Pareto-improving with respect to the status quo where loans are fully absent from the picture.

However, people live in communities where social norms apply, and changes brought to a few individuals can have serious consequences for the whole society.[8] A typical example stems from an AIDS prevention intervention that implies testing all the participants for HIV. In this case, and in many others, the study can have major destabilizing effects. This is particularly relevant in contexts where social norms are disrupted (Morvant-Roux et al. 2014). Backlashes are typically observed in the context of interventions intended to favor women's empowerment in patriarchal societies (Schuler et al. 2018). As theorized by Rabin (1993), economic fairness has welfare implications. People can find it acceptable to get nothing when it is everybody's fate but react negatively if they keep getting nothing while others are randomly, and thus "unjustly," rewarded. Moreover, if the tested treatment is expected to make such a little difference, then maybe it does not deserve to be investigated with such a heavy (and expensive) experimental design. Using RCTs implicitly supposes that the stakes are high enough to deploy a sophisticated and costly experimental design. Put boldly, failing to account for—at least some sort of—equipoise implies that the study either is a waste of money or causes "moral discomfort" (Baele 2013: 4).

Unlike medical doctors, development economists are not used to seeing ethics interfere with their research methods. Generally, ethics is not mentioned as a concern by the "randomistas," let alone as a goal in experimental design (Barrett and Carter 2010). Their objectives are elsewhere, typically in ambitious policy implications to "solve poverty" (Karlan and Appel 2011) and in testing economic theories. Banerjee and Duflo (2009: 156) view experiments as "a powerful tool for testing theories." Both types of concerns deviate from that of medical investigators.

Arguably, policy recommendations are closer to the physician's interest for the situation of her patient. There is, however, a notable nuance since economic policy is intended as a general treatment for a large group of people, some of which are not even seeking treatment. The take-up problem frequently met by randomistas occurs when participants are not interested in the treatment. The investigator's typical response is triggering take-up with some sort of encouragement (White

[8] To mitigate this problem, randomization is sometimes clustered at the community level. Cluster-RCTs can still raise the concern of equipoise. For example, in some RCTs exploring the distribution of insecticide-impregnated bed nets, there is only one group that receives them for free (Müller et al., 2008; Tarozzi et al., 2014).

2013), which can in turn have unwanted consequences on the empirical results. By contrast, medical RCTs use local field investigators to assess the condition of the patients to be enrolled in the study. This procedure prevents researchers from including non-sick participants, both in the treated group and in the control group. For economists, proceeding in the same way without doing any harm would mean identifying the needs of all the participants and so building two types of options (possibly including some compensation) of ex ante equal interest to be provided randomly to the treated and control groups. Hence equipoise.

The contrast between the objectives of policy vs. theory is reminiscent of what Weijer, Glass and Shapiro (2000) present as a UK vs. US controversy, or more broadly, as the science vs. patient trade-off. Yet, in experimental economics there is no consensus on a middle-of-the-road option that would offer to the control group the equivalent of the "standard treatment plus placebo" in order not to harm the individuals who do not receive the experimental treatment. In fact, most economic RCTs go in the opposite direction by testing treatments about which the expectations of positive outcomes are the highest. This is the reason why the microfinance community was so disappointed to see that the results of the related RCTs were so modest (Duvendack et al. 2011). If the experiments were constrained by equipoise, this community would probably have been less disappointed following the mitigated outcomes of RCTs. But paradoxically, the modest impacts captured by these RCTs could also be presented as ex post rationalization that some kind of equipoise was met. A key issue in this field of research stems however from the fact that the outcomes checked vary across publications, leading to testing disparate impacts on subjects' well-being. For instance, Attanasio et al. (2015) use a large set of outcomes including, e.a., increase in entrepreneurship, schooling, consumption, and repayment rates. This outcome heterogeneity adds another layer of complexity to the implementation of equipoise in social sciences.[9]

The short shrift given to ethics in economic RCT conversations echoes the past lack of reactions following unethical medical experimentation on disadvantaged populations. In the Tuskegee Study (1932–1972) sponsored by the U.S. Public Health Service, the group of interest was made up of poor African-American men with untreated syphilis (Caplan 2001). The justification provided by the experimenters to give no treatment to the control group was that these poor men would not afford the treatment anyway (Angell 1997). This argument may sound familiar to those who have questioned randomistas about the unfairness toward the members of their control groups. Interestingly, the Tuskegee Study stopped following an intervention of the media—namely the *Washington Star* and the *New York Times*—that embarrassed the Nixon administration (Angell 1997).

[9] The issue of multiple outcomes goes beyond the microfinance literature. A recent study by Schilbach (2019) on the impact of commitment devices on alcohol consumption among cycle-rickshaw drivers in India uses outcomes such as alcohol consumption, daytime sobriety, productivity, earnings, and savings. In any case, randomly assigning sobriety incentives is ethically questionable.

Lurie and Wolfe (1997) point out the different study designs used in the U.S. and in developing countries[10] in medical RCTs testing new drugs intended to save the life of infants born to HIV-infected women. The authors oppose the situation of the two trials performed in the U.S., in which the study groups had access to antiretroviral drugs, to those that took place in developing countries, where most patients had no such access, likely for financial reasons. Lurie and Wolfe (1997) also report the anecdote of a Harvard researcher who applied to the U.S. National Institutes of Health (NIH) to get funding for an ethically well-designed—with an actively treated control group—study in Thailand and received from the NIH the cost-reducing recommendation to run a placebo-controlled trial instead. It is only after the director of Harvard's human subjects committee replied that, for such a situation, placebo-controlled trials were unethical that the NIH accepted the argument. Beyond anecdotal evidence, Petryna (2009) observes that the share of medical RCTs organized in emerging markets[11] increased from 10 percent in 1991 to 40 percent in 2005. She questions the exploitative character of offshoring clinical trials, which can be used by pharmaceutical companies to encourage doctors in emerging countries to prescribe high-cost medicines, and so undermine the delivery of affordable treatments. Evidently, the bulk of experimental economists have no such profit-oriented motivations. Yet RCTs are costly to implement and so divert money away from other, often less-consuming, experimental designs. Moreover, the history of medical trials in developing countries pinpoints the reputational threats associated with experiments involving poor populations. These populations are easy to exploit because they are mostly unaware of their rights to full disclosure about the experimental design, and to subsequent informed consent or denial.

Another notable difference between medical and economic studies relates to the scope—rather than the nature—of the impact sought, since the economic perturbations brought to existing structures and traditions are believed to be minor compared to drug testing. In addition, most RCTs performed by economists in developing countries seek evidence that relates one way or another to the purpose of aiding these countries. These studies cannot be suspected of having a commercial agenda or using developing countries as a lab for treatments intended for developed economies. Yet, RCTs without equipoise allocate supposedly favorable situations to randomly chosen individuals. Following an argument often used when experimental drugs are rationed, one would expect that the treatment would be offered to those who are the most in need. Along this line of thought, Baele (2013: 19) claims that "randomization also violates the prioritarian moral principle of ensuring a certain level of well-being to the worse-off subpopulation before thinking of either maximising in absolute terms the wealth of

[10] Ivory Coast, Uganda, Tanzania, South Africa, Malawi, Thailand, Ethiopia, Burkina Faso, Zimbabwe, Kenya, and the Dominican Republic.

[11] The author cites several examples involving Eastern Europe and Brazil.

the population (consequentialist version) or ensuring individual freedom (liberal version)." Accordingly, even benign economic treatments are no reason for deviating from the ethical requirement of equipoise toward the control group.

If the economic profession finds the equipoise requirement suitable, its implementation would lead to constraining study design and imposing that the treated and control groups get qualitatively similar outcomes, where similarity must be backed by expert opinion. However, the funders' preference for promising impact can put pressure in the direction opposite to equipoise (Ravallion, Chapter 1). Overall, the practical implementation of this procedure can be tedious, but at least it is expected to exclude blatantly unethical behaviors.

In addition, the medical examples described in this section reveal that even in medical sciences where the Declaration of Helsinki and the equipoise requirement are supposed to act as moral compasses, ethical relativism dies hard. Researchers performing RCTs in developing countries tend to lower the standards of care based on the financial argument that most people—and/or their governments—cannot pay for the best treatments. Dismissing such arguments, the editorial article by Marcia Angell (1997: 848) announced that the prestigious *New England Journal of Medicine* has decided not to publish studies reporting "unethical research, regardless of their scientific merit." Will economic journals follow the same path? At the moment, there is no indication of such a move.

10.5 Conclusion

The time has come to ask RCT development economists, and other randomistas, to confront the ethical consequences of their doings with the ultimate goals of the trials they launch. Like in the medical community in the 1980s, more and more scholars are questioning the "gold standard" paradigm (Cartwright 2007; Deaton 2010a; Bédécarrats, Guérin, and Roubaud 2019) along several dimensions. Even though RCTs can affect lives significantly, criticisms based on their ethical dimension are so far among the least vocal ones. Arguably, this is due to the underlying good intentions, such as helping poor people address malaria prevention (Cohen and Dupas 2010) and sanitation threads (Duflo et al. 2015). At the same time, as Baldassarri and Abascal (2017: 62) put it, "field experimenters 'play God,' intervening in people's lives in consequential ways." Often, previous field knowledge makes negative outcomes predictable with some degree of confidence. In such cases, the lack of equipoise associated with the specific risks to which disadvantaged people are exposed result from either the indifference or the insufficient field experience of the experimenters. From an ethical standpoint, this situation is worse than that of studies randomly assigning presumably favorable treatments. Yet, both can have long-lasting effects not only at the individual level, but also on interpersonal relationships within communities.

On the other hand, arguments focusing on the unfairness entailed by the experimental design neglect the potential long-term benefits brought by the novel knowledge that RCTs are built to deliver. Even in the medical sphere, some authors criticize the very notion of equipoise or propose to amend it (Fries and Krishnan 2004; Ubel and Silbergleit 2011). The main concern relates to failing to account for the future benefits that are disregarded when dealing with the therapeutic obligation of physicians toward their patients. Veatch (2007: 182) claims that "it is not anyone's equipoise that is morally critical; it is whether the potential subjects consent to be randomized without being unduly coerced, manipulated, or exploited." Miller and Brody (2007: 153) argue "that the ethical principles governing medical therapy are different from those governing clinical research." The authors emphasize the importance of the returns expected from the study as a legitimate reason to deviate from the requirement of equipoise. This logic corresponds to accepting to sacrifice the welfare of some individuals (typically those in the control group) for the greater good of society and future generations. It is however mitigated by the double blindness that makes the sacrifice probabilistic rather than certain, and so spreads it over all the study participants. Economic RCTs mostly use no blindness, which makes it more difficult to refer to this convenient excuse. In any case, scientific studies treating unfairly or inflicting sacrifices to people, and especially to those already disadvantaged, will always be ethically questionable. And since morally debatable studies are preferentially organized in countries where legal protection is weak, the issue of imposing—at least some form of—equipoise to economic experimentation should go beyond the rhetoric.

Acknowledgements

The authors thank Cécile Abramowicz, Britta Augsburg, Marie Brière, Andres Garchitorena, Marek Hudon, Marc Labie, Jonathan Morduch, Tim Ogden, Martin Ravallion, the participants to the AFD Workshop *Randomized Control Trials in the Field of Development: The Gold Standard Revisited* (Paris, March 2019), and the three editors of the book, Florent Bédécarrats, Isabelle Guérin, and François Roubaud, for valuable comments and discussions.

11
Using Priors in Experimental Design
How Much Are We Leaving on the Table?

Eva Vivalt

11.1 Introduction

There has been much debate about the relative merits of RCTs. In this chapter, I abstract from the larger debate to focus on one relevant narrow and under-considered issue: to what extent leveraging priors can improve study design. By priors, I mean prior beliefs about the effects of a particular program held by decision-makers. For example, some policy-makers may believe an unconditional cash transfer program is very likely to have large impacts on educational outcomes. If the decision-making process and the priors were known, it would enable researchers to better target their impact evaluations. In some cases, there may be many treatment arms that could be tested, and priors could inform which arms are run and evaluated. Alternatively, the treatment arms may be fixed but there could be some question of which outcome measures to prioritize. In particular, a common decision that researchers face is which outcomes to include in midline or endline surveys, given time constraints. Those outcomes gathered more frequently will be better-powered, all else equal, so decision-makers' priors could inform which outcomes to gather more frequently. One can also imagine that depending on decision-makers' priors, the sample size of the study may need to be larger or smaller in order for the study to provide convincing evidence, so use of priors could make an impact evaluation more efficient at informing policy.

Both RCTs and studies leveraging quasi-experimental methods (henceforth "non-RCTs") could in principle profitably leverage priors. However, these methods also interact with the use of priors. To the extent to which RCTs can be thought of as motivated by the desire to convince a maximally adversarial decision-maker (Banerjee et al. 2017a), leveraging a *specific* set of priors seems philosophically more aligned with non-RCTs, i.e., we may worry less about a maximally adversarial audience if the study is set up to convince a specific audience with known priors. Further, only non-RCTs can deterministically assign participants to treatment groups, which is necessary for some of the potential gains from leveraging priors. On the other hand, there may be some justifiable

Eva Vivalt, *Using Priors in Experimental Design: How Much Are We Leaving on the Table?* In: *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0013

skepticism of the results of non-RCTs compared to RCTs on the grounds that it may be easier for researchers to consciously or subconsciously engage in specification searching in non-RCTs. Specification searching, or "p-hacking," occurs when researchers conduct many statistical tests (such as by running many regressions with different control variables) and preferentially report the results of those that are significant. This leads to incorrect inferences. More specification searching has repeatedly been observed in non-RCTs using classical significance tests (Brodeur et al. 2016, 2018; Vivalt 2019).[1] As collecting and leveraging priors is still a new approach in economics, it could be more appealing to do for RCTs, as the choices made in the process would be more transparent. The rest of this chapter will elaborate on these points, after further describing what is meant by prior beliefs and how they may be elicited.

11.2 The Elicitation and Use of Priors

A growing number of researchers have been collecting ex ante priors as to the effects that their studies will find. For example, a team doing an impact evaluation of a conditional cash transfer program, in which households are given money in exchange for sending their school-aged children to school, might ask others to provide a best guess as to the effect the program would have on enrollment rates. As part of this process, the researcher team would first describe the program and its context in great detail so as to encourage more accurate guesses.

There are several reasons why one might want to collect these prior beliefs. First, these forecasts can be interesting in and of themselves, as one can learn whether certain subgroups of respondents make more accurate predictions. For example, some work suggests that policy-makers tend to have more optimistic beliefs than researchers (Casey et al. 2018; Vivalt and Coville 2016). Over time, we may learn about the conditions under which individuals make better forecasts or identify individuals who are better at making. We could also learn how to "de-bias" the forecasts as much as possible through modeling or a machine learning approach. At the end of the process, we are left with a potentially valuable output: forecasts with some informational content.

These forecasts can be important for policy-making as we are never able to run as many studies as we would like. In the absence of evidence from academic studies, de-biased forecasts can help policy-makers decide which interventions to

[1] We normally think of specification searching as something that occurs when using classical (frequentist) tests, e.g. when testing whether some relationship appears significant at the 5 percent level. However, it is possible that even in a Bayesian setting, if researchers knew the prior beliefs of decision-makers they could engage in specification searching to either support or undermine the decision-makers' beliefs and affect policy decisions. Therefore, even to a Bayesian, RCTs may appear more credible.

run. Note that we do not need the forecasts to always be accurate to be valuable for this purpose. The forecasts need only be correlated with the interventions' effects to be useful on net, though mistakes would still be made.

More mundanely, individual research teams have private incentives to collect ex ante prior beliefs: presently, there may be a publication benefit to doing so. In particular, academic journals often put a premium on results that are statistically significant. By collecting ex ante priors, researchers have some protection in the state of the world in which they obtain "null" results (i.e. in which they find the program had zero effect). In such cases, the priors will sometimes allow them to credibly argue that these null results were unexpected and hence still of academic interest.[2]

The rest of this chapter will focus on yet another benefit of collecting ex ante prior beliefs about the effects of interventions: their potential benefits in informing study design.

11.3 The Use of Priors in Study Design

Priors can inform study design through several mechanisms. First, one can change one's allocation of a potential sample to different treatment groups or collect different outcome variables depending on the information value of each allocation. Yet priors could also improve a study's design through deterministic assignment, a fact that has been known for a very long time (see, for example, debates between Fisher, 1960, and Gosset, 1937). For background, it may be helpful to review an example from Banerjee et al. (2017a) in which an educational vouchers experiment is planned to help a school superintendant decide whether or not to use vouchers. The school superintendant believes that whether a student is from a rich or poor family is a major determinant of academic success, but the superintendant is also open to the idea that school quality may be very important, such that even a poor student could learn more at a private school. The superintendant is allowed to assign one student to a private school. From a classical perspective, it is impossible to learn anything meaningful from an experiment with one observation. However, to a Bayesian, something could still be learned. Namely, if a poor child were assigned to the private school and subsequently did better on standardized tests than the superintendant's prior beliefs as to how well a poor child might score, that would be informative and the superintendant should update their beliefs.

[2] In the long run, if collecting priors takes off, one might imagine that judging the novelty and importance of research results using prior beliefs would supplant comparisons against the null of zero effect and publication bias would shift to those studies with "novel" results.

The superintendant's priors play an important role here. Due to their priors, assigning a rich child to the private school would not be as informative as assigning a poor child. Likewise, if a slot in a public school became available, it would not make sense to assign a poor child to the public school from a value of information perspective. What these examples show is that a Bayesian should assign subjects to treatment and control groups deterministically, rather than randomly. Similar arguments have been made elsewhere (Kasy 2016).

Banerjee et al. (2017a) go on to show that in order to convince an adversarial audience, randomization can help. For example, suppose that rather than there being one superintendant with a single set of beliefs, there were a pool of possible superintendants with the full range of possible beliefs about how academic success depends on whether a student is rich or poor and whether they attend a private or public school.[3] Banerjee et al. (2017a) ask us to consider the case in which, after an experimental design is chosen, the superintendant with the prior that maximizes the chance the wrong policy will be implemented is chosen from this pool of possible superintendants. In this "adversarial" world, there are no longer gains to allocating a poor child to a private school and a rich child to a public school. Likewise, if a decision-maker did not trust their own prior and was ambiguity averse, randomization could help as a mixed strategy.

Banerjee et al.'s (2017a) insights are excellent and explain why researchers seeking to convince referees prefer RCTs—referees are certainly a maximally adversarial audience! Their arguments also explain why firms running experiments for internal use, often with small sample sizes, are less likely to randomize: with smaller sample sizes available, the gains in statistical power from deterministic assignment are larger, and they may design their evaluation in order to make a specific decision given some particular priors. However, policy-makers seem to fall somewhere between these extremes. They are not likely to be maximally adversarial but may instead have priors that lie within a restricted range. Yet policy-makers are also not quite like firms, as they may be able to leverage larger sample sizes for their trials and as there may be some uncertainty over the exact set of priors that will be relevant in the future, e.g. if there is an election or another source of staff turnover before the results of the evaluation will become known. Policy-makers may also more often intend the evaluation to provide a public good for other policy-makers. They may additionally worry about convincing their constituents, who may hold a wide range of prior beliefs, about the merits of the social program. To the extent to which others' priors are unknown or adversarial, RCTs may be preferable to an ambiguity averse policy-maker.

An argument could also be made that researchers should focus on RCTs if biases are likely to creep into quasi-experimental studies. Brodeur et al.

[3] For example, there could be turnover in school superintendents, resulting in some amount of variation in beliefs.

(2016, 2018) found RCTs and laboratory experiments had fewer results just above, as opposed to just below, the classical 5 percent threshold for significance than other studies;[4] using another data set, Vivalt (2019) found 17 percent fewer RCTs fell just above the threshold than non-RCTs. This is convincing evidence that RCTs currently exhibit fewer signs of specification searching than non-RCTs when classical significance tests are considered. A Bayesian would not care about classical significance tests; however, there are other ways in which results may be distorted. In particular, specification searching that inflates the significance of classical tests will also tend to have the effect of exaggerating the magnitude of the estimated effect, so if research is being done both for publication, where the classical tests matter, and to inform a policy decision, where policy-makers might be Bayesian, one may still question the magnitude of the reported effects.

RCTs may exhibit less classical specification searching for two types of reasons, each of which has different implications for Bayesians. First, researchers may have fewer incentives to engage in specification searching if their studies have an easier time being published. Second, there may be something about RCTs that intrinsically makes specification searching less likely, such as researchers more often using pre-analysis plans that increase the difficulty of specification searching; it potentially being harder to justify actions like including various control variables since in principle randomization should lead to covariate balance in large samples; or RCTs carrying connotations of rigor that nudge authors to not engage in bad practices. The first issue may not apply in the context of Bayesian decision-making, though if researchers wanted to influence a policy decision they may have similar incentives to distort results. To the extent to which the second type of issue is a factor driving specification searching in the classical case, it would remain relevant in the Bayesian case.

It might be sensible to assume that any specification searching would merely exaggerate effects, rather than changing the sign of an effect, and so if one were Bayesian the correct conclusion would not be to completely distrust non-RCT results but to perhaps shrink the estimates towards zero by some typical "exaggeration factor."[5] If one believes non-RCTs to be biased, one may ironically be more adversarial towards their results. Yet these biases should not mean wholesale abandonment of non-RCTs, either. Better norms and commitment devices such as pre-analysis plans could help to address this issue.

The above discussion has focused on one particular type of bias that may affect a study's results. The issue of whether non-RCTs are more biased than RCTs is

[4] Brodeur et al. (2018), however, did find that one of the kinds of quasi-experimental designs they considered, regression discontinuity designs, did better than RCTs.

[5] For example, Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) introduce "Type M" errors (for "magnitude"), which could describe how much a result is likely to be exaggerated.

more complicated than the above discussion suggests. In particular, in other chapters it is argued that RCTs may be done in selected locations, causing site selection bias, or that they may have smaller samples than non-RCTs, causing their results to have larger confidence intervals and also resulting in predictions based on the results to have larger mean squared errors (due to the bias-variance trade-off). I abstract from these potential issues for two reasons. First, in other work leveraging a sample of 635 studies of 20 interventions in international development, I fail to reject the null hypothesis that the effect sizes estimated by RCTs and non-RCTs are the same (Vivalt forthcoming). Second, there are many situations in which the costs of RCTs and non-RCTs are identical, so that it is not necessarily the case that RCTs will have smaller sample sizes. Nonetheless, the bias-variance tradeoff is an extremely under-appreciated issue in economics. In an ideal world, a policy-maker should care not just about whether the point estimate is biased but also about the precision of the estimates and the total prediction error. There is some evidence that policy-makers instead suffer from "variance neglect," partially misunderstanding or ignoring confidence intervals (Vivalt and Coville 2016).

11.4 Back-of-the-Envelope Estimates of Benefits of Using Priors

The question remains: what are the benefits of using priors in study design? I conduct some simple simulations that provide back-of-the-envelope estimates. In this section, I consider potential benefits that are held in common across RCTs and non-RCTs to emphasize the point that leveraging prior beliefs, like many other important methodological issues, can be important for both RCTs and non-RCTs.

Suppose there are two interventions—such as a conditional cash transfer program and a school meals program—and a policy-maker is deciding which to implement in order to improve enrollment rates in school. In order to make this decision, the policy-maker asks a researcher to plan an impact evaluation of a pilot of one of the two interventions. If the researcher does not know the priors of the policy-maker, they will not know which intervention to do the impact evaluation on and could pick the wrong intervention to study. For example, suppose the policy-maker was very uncertain about the effects of the school meals program but quite certain about the effects of a conditional cash transfer program. Then it would make sense to do the impact evaluation on the school meals program, as the policy-maker would be unlikely to change their mind about the effects of the conditional cash transfer program.

I generate back-of-the-envelope estimates of the benefits of considering priors using the following algorithm: first, I specify a prior for each intervention. For simplicity, I assume normally distributed priors, so this requires simply

specifying a mean and standard deviation. I assume that the standard deviation of the outcome variable for each intervention is the same and that in the absence of using priors each intervention would be equally likely to be selected for the impact evaluation. I then determine the value of information, from the policy-maker's perspective, for doing an impact evaluation of each intervention given the priors that were drawn. The value of information is constructed as the probability an impact evaluation would change the decision that was made, with the policy-maker always preferring the intervention with the higher mean posterior, multiplied by the expected benefit of making that decision, i.e. the difference between the true means of the two interventions. Again, it should be emphasized this calculation is based on the policy-maker's beliefs, which could be incorrect. I do these calculations making different assumptions about the precision of the impact evaluation's results. This value of information is used to determine which program would be selected for study. I then calculate the posteriors the policy-maker would hold after doing a study of that program, assuming certain true mean values for each program.[6] Finally, I estimate the value of leveraging priors as the difference in outcomes of the programs ultimately selected for post-evaluation implementation if one were to always pick the treatment arm that has the higher value of information as opposed to picking the right intervention to study with 50/50 probability.

Table 11.1 summarizes results for several confidence intervals, priors, and true mean values, with "A" representing the mean impact of CCTs in this example and "B" representing the mean impact of school meals programs. The estimates are in their raw units—percentage point increases in enrollment rates. The table can be read as follows: the first column provides the assumed mean prior for program "A," with the prior mean for program "B" assumed to be zero; the second and third columns provide hypothetical standard deviations for the priors for programs "A" and "B," respectively; the fourth column provides various hypothetical values for the true impact of program "B," where program "A" is assumed to have zero impact; the last three columns present the simulated benefits of using priors to determine which program to evaluate in terms of the increased impact of the program selected post-evaluation for each of several confidence interval widths. The confidence interval widths here span from the top to bottom of the 95 percent range. For example, for a confidence interval of 0.1 the standard error would be equal to 0.1/(2*1.96). I assume that decision-makers are Bayesian, so that when they determine which program to pursue post-evaluation they do so by appropriately combining their priors with the new evidence provided by the impact evaluation.

[6] I restrict attention to the mean, though other parts of the distribution of outcomes may also be of interest.

Table 11.1 Estimates of benefits from considering priors

Priors				Benefits to using priors, for various confidence intervals		
Mean for A	**SD for A**	**SD for B**	**Mean B**	**0.01**	**0.1**	**1**
0.1	0.1	2.0	1	0.5	0.5	0.5
0.1	1.0	2.0	1	0.5	0.5	0.5
0.1	5.0	2.0	1	−0.5	−0.5	−0.5
0.5	0.1	2.0	1	0.5	0.5	0.5
0.5	1.0	2.0	1	0.5	0.5	0.5
0.5	5.0	2.0	1	−0.5	−0.5	−0.5
1.0	0.1	2.0	1	0.0	0.0	0.0
1.0	1.0	2.0	1	0.0	0.0	0.0
1.0	5.0	2.0	1	0.0	0.0	0.0
5.0	0.1	2.0	1	0.0	0.0	0.0
5.0	1.0	2.0	1	0.0	0.0	0.0
5.0	5.0	2.0	1	0.0	0.0	0.0
0.1	0.1	2.0	5	2.5	2.5	2.5
0.1	1.0	2.0	5	2.5	2.5	2.5
0.1	5.0	2.0	5	−2.5	−2.5	−2.5
0.5	0.1	2.0	5	2.5	2.5	2.5
0.5	1.0	2.0	5	2.5	2.5	2.5
0.5	5.0	2.0	5	−2.5	−2.5	−2.5
1.0	0.1	2.0	5	2.5	2.5	2.5
1.0	1.0	2.0	5	2.5	2.5	2.5
1.0	5.0	2.0	5	−2.5	−2.5	−2.5
5.0	0.1	2.0	5	0.0	0.0	0.0
5.0	1.0	2.0	5	0.0	0.0	0.0
5.0	5.0	2.0	5	0.0	0.0	0.0
0.1	0.1	2.0	10	5.0	5.0	5.0
0.1	1.0	2.0	10	5.0	5.0	5.0
0.1	5.0	2.0	10	−5.0	−5.0	−5.0
0.5	0.1	2.0	10	5.0	5.0	5.0
0.5	1.0	2.0	10	5.0	5.0	5.0
0.5	5.0	2.0	10	−5.0	−5.0	−5.0
1.0	0.1	2.0	10	5.0	5.0	5.0
1.0	1.0	2.0	10	5.0	5.0	5.0
1.0	5.0	2.0	10	−5.0	−5.0	−5.0
5.0	0.1	2.0	10	5.0	5.0	5.0
5.0	1.0	2.0	10	5.0	5.0	5.0
5.0	5.0	2.0	10	−5.0	−5.0	−5.0

Source: Author.

These simulations suggest the benefits can be quite large. For the values I selected, the benefits were as large as the ultimately-selected program having a 5 percentage point larger effect, i.e. with it increasing enrollment rates by 5 percentage points rather than by 0 percentage points. However, the benefits greatly depend on the priors and, of course, on the true means of each program.

Table 11.2 Estimates of benefits for different prior values

Priors				Benefits to using priors, for various confidence intervals		
Mean for A	**SD for A**	**SD for B**	**Mean B**	**0.01**	**0.1**	**1**
5.1	0.1	2.0	1	−0.5	−0.5	−0.5
5.1	1.0	2.0	1	−0.5	−0.5	−0.5
5.1	5.0	2.0	1	0.5	0.5	0.5
5.5	0.1	2.0	1	−0.5	−0.5	−0.5
5.5	1.0	2.0	1	−0.5	−0.5	−0.5
5.5	5.0	2.0	1	0.5	0.5	0.5
6.0	0.1	2.0	1	−0.5	−0.5	0.0
6.0	1.0	2.0	1	−0.5	−0.5	−0.5
6.0	5.0	2.0	1	0.5	0.5	0.5
10.0	0.1	2.0	1	−0.5	−0.5	0.0
10.0	1.0	2.0	1	−0.5	−0.5	−0.5
10.0	5.0	2.0	1	0.5	0.5	0.5
5.1	0.1	2.0	5	−2.5	−2.5	−2.5
5.1	1.0	2.0	5	−2.5	−2.5	−2.5
5.1	5.0	2.0	5	2.5	2.5	2.5
5.5	0.1	2.0	5	−2.5	−2.5	−2.5
5.5	1.0	2.0	5	−2.5	−2.5	−2.5
5.5	5.0	2.0	5	2.5	2.5	2.5
6.0	0.1	2.0	5	−2.5	−2.5	0.0
6.0	1.0	2.0	5	−2.5	−2.5	−2.5
6.0	5.0	2.0	5	2.5	2.5	2.5
10.0	0.1	2.0	5	−2.5	−2.5	0.0
10.0	1.0	2.0	5	−2.5	−2.5	−2.5
10.0	5.0	2.0	5	2.5	2.5	2.5
5.1	0.1	2.0	10	0.0	0.0	0.0
5.1	1.0	2.0	10	0.0	0.0	0.0
5.1	5.0	2.0	10	0.0	0.0	0.0
5.5	0.1	2.0	10	0.0	0.0	0.0
5.5	1.0	2.0	10	0.0	0.0	0.0
5.5	5.0	2.0	10	0.0	0.0	0.0
6.0	0.1	2.0	10	0.0	0.0	5.0
6.0	1.0	2.0	10	0.0	0.0	0.0
6.0	5.0	2.0	10	0.0	0.0	0.0
10.0	0.1	2.0	10	−5.0	−5.0	0.0
10.0	1.0	2.0	10	−5.0	−5.0	−5.0
10.0	5.0	2.0	10	5.0	5.0	5.0

Source: Author.

In this example, leveraging prior beliefs could result in an expected benefit of half the difference in true impacts of the two potential programs. This makes sense: leveraging priors causes individuals to learn information that makes them switch which program they would have implemented when they would have only learned that information half the time in the absence of leveraging priors. On the

other hand, it is not the case that leveraging prior beliefs will always help. For some priors, seeing evidence from an impact evaluation would not be enough to nudge decision-makers into making a better decision about which to implement post-evaluation. If they have more uncertainty about a particular program and yet would not change their views given the impact evaluation results for that program (such as if the evaluation were low-powered), using priors to determine which to evaluate could still result in a decision-maker picking the worse program for implementation post-evaluation.

Whether leveraging priors is helpful thus depends in great deal on the priors themselves. Table 11.2 presents simulations using the same inputs except for the prior mean for program "A" and the prior mean for program "B" each being 5 percentage points higher than they were in the simulations presented in Table 11.1. Now, rather than helping for most of the values chosen, leveraging priors hurts most of the time. The values chosen for the prior distributions used to create this table could reflect over-optimism about programs' results, which has been observed in several studies (Casey et al. 2018; Vivalt and Coville 2016). However, if these biases were systematic and could be predicted, they could be corrected for and not have deleterious effects. For still other prior distributions, leveraging priors may neither help nor hurt but simply have no effect in the decision-making process, such as in cases in which a decision-maker is over-confident in their beliefs. Again, in the long run one might expect decision-makers to be sophisticated and able to at least partially correct for their over-optimism or over-confidence. This would require much change to decision-making processes, but there are already indications that eliciting, modeling, and using priors is catching on (e.g. the DARPA SCORE project, DellaVigna and Pope 2018a, 2018b, and DellaVigna, Pope and Vivalt 2019).

11.5 Conclusion

One may wish to conduct a similar exercise using real priors for a variety of interventions and, moreover, to explore deterministic assignment. While I have collected a variety of priors data in other projects (Vivalt and Coville 2016; Coville and Vivalt 2017), they are not trivial to use for this purpose because typically interventions aim to affect different outcome variables (e.g. enrollment rates vs. diarrhea prevalence), and in order to make comparisons across different outcome variables one needs to be able to assess the relative value of the outcomes, a task far beyond the scope of this chapter.

Instead, I offer some remarks. My focus throughout has been on the commonalities between RCTs and non-RCTs, though I have also noted differences where they exist. I do not focus much on these differences, though, because I expect the value of information benefits of deterministic assignment will not be pivotal in the decision of which to use. In particular, sometimes an RCT is impossible and

only quasi-experimental methods can be used—in that case there is no real choice to be made. Sometimes there is only political will for an RCT, so the question of which to use is again moot. While in principle, I would argue that one should consider methods on a case-by-case basis, in practice this does not seem like a choice researchers are frequently able to make.

While the debate pits RCTs against non-RCTs, other issues in experimental design also seem potentially more important and more neglected. First, remarkably few economic researchers or policy-makers appear to be Bayesian, which can lead to prioritizing significant results with small effects over less certain but potentially more effective programs. Second, both RCTs and non-RCTs are generally very limited in what they study and exclude most indirect effects. For example, suppose there is an education program that may affect later-life income or health. It is rare that anyone will return to study these effects years later. There may also be spillovers to institutions or subsequent generations. Most of the time, these possibilities are ignored. There are reasons for this: it is very hard to capture all the relevant effects. Nonetheless, there are methods that could be used in conjunction with either an RCT or a non-RCT, such as Athey et al.'s "surrogacy score" approach to estimating long-run outcomes (2016), that are as yet neglected by both RCTs and non-RCTs. Decision-making processes are also far from ideal, and relatively few people thoughtfully consider evidence from either RCTs or non-RCTs, let alone unbiasedly update based on them. Further, much more work is needed to determine how to best elicit and aggregate forecasts so they can be most profitably used. Finally, researchers fail to coordinate in ways that make it hard to synthesize knowledge across multiple studies, such as by not sharing outcome variables in common. Less than 10 percent of the studies in AidGrade's data set of impact evaluation results in international development made their underlying micro-data publicly available (Vivalt forthcoming). Ironically, while RCT critics accuse RCT proponents of missing the bigger picture, they may be making a similar mistake when there are potentially more important battles to be fought.

12

Epilogue: Randomization and Social Policy Evaluation Revisited

James J. Heckman

12.0 Preamble

This chapter updates my published 1992 paper, "Randomization and Social Policy Evaluation"[1] and places it in the context of the research that followed. The paper is still relevant for understanding the fundamental nature of experiments and what can be learned from even "ideal" experiments with no attrition, non-response, and stratification on the outcome variables of interest. It is worth revisiting in light of the continuing controversies surrounding the role of randomization in development economics. The conceptual points made here have not been addressed in the literature, even though many issues of implementation have.

This preamble provides some perspective on the history of field experiments and the origins of the experimental movement in economics. The history of field experimentation in economics since 1965 can be classified into two eras: (1) The early wave that used experiments to settle important policy debates where non-experimental evidence was ambiguous; and (2) the revival of experimentation in development economics that culminated in the 2019 Nobel Prize in Economics. Each era has been marked by a near-religious zeal for the methodology of randomized control trials (RCTs). Accordingly, I name both the eras, "Great Awakenings," in honor of two religious revivals that shaped Protestant churches in North America in the eighteenth and nineteenth centuries, and in recognition of the zeal for methodological purity in both eras in economics.

The First Great Awakening arose in the push to evaluate the manpower, education, and health programs launched by Lyndon Johnson's War on Poverty. The Second Great Awakening came in development economics in the wake of a variety of micro programs targeted to less-developed countries funded by influential NGOs, billionaires, and various international institutions. Few of the hard lessons learned about the limitations of social experiments from the First Great Awakening are acknowledged by the economists promoting the Second Awakening. The

[1] See Heckman (1992).

James J. Heckman, *Epilogue: Randomization and Social Policy Evaluation Revisited In: Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by: Florent Bédécarrats, Isabelle Guérin, and François Roubaud, Oxford University Press (2020).
DOI: 10.1093/oso/9780198865360.003.0014

career incentives of the new generation argue against examining and citing the contributions and lessons of the First Awakening, which ended in substantial qualification of the alleged claim of "transparent results" and eventual decline in the uncritical enthusiasm for RCTs. The Second Awakening will likely suffer the same fate.

12.0.1 The First Awakening

Long before randomization became de rigueur in the field of development in the First Great Awakening, it was advocated for evaluating a variety of social programs, educational interventions, workforce training programs, and welfare reforms.

In the First Wave, leading evaluation firms, such as Westat, Mathematica, SRI, Abt Associates, and MDRC, addressed the mandate of the Office of Economic Opportunity (OEO) that administered Lyndon Johnson's War on Poverty to evaluate a raft of newly launched social programs. The emphasis on evaluation percolated across many U.S. federal agencies.

This early thrust for evaluation led to the collection of novel panel micro data sets that continue to guide understanding of society and are now widely emulated around the world. The First Awakening also fostered new methodologies to analyze the serious problems that plagued the experiments conducted in the First Wave.

The first wide-scale use of randomization in economics was in evaluating Negative Income Tax (NIT) programs. These programs were proposed by Milton Friedman[2] and others as an alternative to the cumbersome welfare transfer programs of the day that heavily taxed low-income workers by substantially reducing benefits for each dollar earned. The NIT was designed to replace the patchwork welfare system of the 1960s by giving a lump sum transfer to the poor and taxing additional earnings at a uniform low rate over the whole income schedule. The policy question was whether imposition of NIT would substantially reduce labor supply. The answer depended on the relative strength of income and substitution effects. Transfers would reduce labor supply through an income effect. The lowered tax rate on earnings would encourage it through a substitution effect. The existing non-experimental estimates of the income and substitution effects ranged all over the place, as documented in the introductory chapter of Cain and Watts (1973).

In the early 1960s, Heather Ross, then a graduate student at MIT, proposed a large-scale randomized trial to gauge the effects of NIT. The Office of Economic Opportunity accepted her proposal and funded it. Many economic consulting firms rose to the challenge. The first NIT experiment was launched in 1968.

[2] See Friedman (2009), reissued.

The early researchers waded into deep waters and sometimes got in over their heads. The initial designs were flawed. Selection bias riddled the studies. Attrition and noncompliance were high. Ironically, analyzing NIT data helped to launch the then nascent field of microeconometrics. The era culminated in John Cogan's testimony before the U.S. Congress,[3] in which he reanalyzed the data from the NIT experiment using the newly developed techniques of microeconometrics. These methods were later recognized by the Nobel Prize Committee in 2000.

Cogan's testimony challenged the "transparent" evidence from the experiment, pointing out a variety of selection biases. He showed negative impacts on labor supply that were substantially larger than the trivial impacts found from the "transparent" experimental comparisons of the mean differences between treatments and controls. At those hearings, Senator Daniel Patrick Moynihan expressed dismay over the low quality of the "transparent" experimental evidence, as revealed by Cogan's analysis, and gratitude to Cogan for presenting an honest report of what the experiment actually demonstrated using non-experimental methods to analyze the flawed experimental data.

12.0.2 The Second Awakening

The Second Wave is in its zenith. The enthusiasm for experimentation has led NGOs, foundations, and governments to mandate its application. Whereas the First Wave was motivated by the desire to address major social questions, the Second Wave has a more methodological focus. It is part and parcel of a professional obsession in the field of economics to obtain "causal effects," even if the effects being identified are without social significance and/or economic meaning.[4] Miniaturist studies became praised as the ideal for rigorous empirical economics. Asking and trying to answer big and important questions was discredited in pursuit of clean answers to small questions of little policy consequence. Indeed, the Nobel Prize Committee in 2019 lauded practitioners in the Second Wave for focusing on "smaller, more manageable problems."[5] The award was for methodological purity and "manageability" rather than for substance.

It is useful to cast the quest of many applied economists marching in the parade of the Second Wave in terms of a traditional regression framework. Let Y be an outcome of interest. Suppose

$$Y = X\beta + D\alpha + U$$

[3] Congress (1978). [4] See Introduction by Deaton, this volume.
[5] Royal Academy of Sciences (2019).

where X is a vector of observed control variables, D is an indicator if treatment is received ($D=1$ if treated, $D=0$ if not), and U is correlated with D. α is "the effect" of treatment controlling for X and U. If we fail to control for X and U, correlational estimates of α are biased, with the sign of the bias determined by the sign of the correlation between U and D controlling for X. Randomization avoids this bias if it is properly conducted.

As in the recent instrumental variables literature, in the Second Awakening eliminating this bias is the paramount issue, usually to the exclusion of asking whether α answers any important question—either in theory or practice. In the First Awakening, that issue was front and center.

The revised paper presented here, and a follow-up paper by Heckman and Smith (1995), were written after the First Wave of enthusiasm for RCTs and before the Second Wave. Both papers are relevant today. The fact that the Second Wave emerged is a tribute either to the bad writing of those papers, or to the demonstrated ability of economists to ignore hard-won lessons from the past, as well as strong career incentives to pour old wine into new bottles and forget its sources. I now turn to the original paper.

12.1 Introduction

This paper considers the benefits and limitations of *randomized* social experimentation as a tool for evaluating social programs.[6] The argument for social experimentation is by now familiar. Available cross-section and time-series data often possess insufficient variability in critical explanatory variables to enable analysts to develop convincing estimates of the impacts of social programs on target outcome variables. By collecting data to induce more variation in the explanatory variables, more precise estimates of policy impacts are possible. In addition, controlled variation in explanatory variables can make endogenous variables exogenous; that is, it can induce independent variation in observed variables relative to unobserved variables. Social experiments induce variation by controlling the way data are collected. Randomization is one way to induce extra variation, but it is by no means the only way or even necessarily the best way to achieve the desired variation.

The original case for social experimentation took as its point of departure the Haavelmo (1944)–Marschak (1953)–Tinbergen (1956) social planning paradigm. Social science knowledge was thought to be sufficiently advanced to be able to

[6] Throughout this paper I refrain from restating familiar arguments about the limitations of social experiments and focus on a problem not treated in the literature on this topic. See Cook and Campbell (1979), the papers in Hausman and Wise (1985a), and the other chapters in this volume for statements on problems of attrition, spillover effects, and so forth.

identify basic behavioral relationships which, when estimated, could be used to evaluate the impacts of a whole host of social programs, none of which had actually been implemented at the time of the evaluation. The "structural equation" approach to social policy evaluation promised to enable analysts to simulate a wide array of counterfactuals that could be the basis for "optimal" social policymaking. The goal of social experimentation, as envisioned by Conlisk and Watts (1969) and Conlisk (1973), was to develop better estimates of the structural equations needed to perform the simulation of counterfactuals.

The original proponents of the experimental method in economics focused on the inability of cross-section studies of labor supply to isolate "income" and "substitution" effects needed to estimate the impact of negative income taxes (NIT) on labor supply. Experiments were designed to induce greater variation in wages and incomes across individuals to afford better estimation of critical policy parameters. The original goal of these experiments was not to evaluate a specific set of NIT programs but to estimate parameters that could be used to assess the impacts of those and many other possible programs.

As the NIT experiments were implemented, their administrators began to expect less from them. Attention focused on evaluations of specific treatment effects actually in place (see Cain 1975). Extrapolating from and interpolation between, the estimated treatment effects took the place of counterfactual policy simulations based on estimated structural parameters as the method of choice for evaluating proposed programs not actually implemented (see Hausman and Wise 1985b).

The recent case for randomized social experiments represents a dramatic retreat from the ambitious program of "optimal" social policy analysis that was never fully embraced by most economists and was not embraced at all by other social scientists. Considerable skepticism had recently been expressed about the value of econometric or statistical methods for estimating the impacts of specific social programs or the parameters of "structural" equations required to stimulate social programs not yet in place. Influential studies by LaLonde (1986) and Fraker and Maynard (1987) convinced many that econometric and statistical methods are incapable of estimating true program impacts from non-randomized data.

Recent advocates of social experiments are more modest in their ambitions than were the original proponents. They propose to use randomization to evaluate programs actually in place (whether ongoing programs or pilot "demonstration" projects) and to avoid invoking the litany of often unconvincing assumptions that underlie "structural" or "econometric" or "statistical" approaches to program evaluations.[7] Their case for randomization is powerfully simple and convincing: randomly assign persons to a program and compare target responses of participants

[7] In an early contribution, Orcutt and Orcutt (1968) suggest this use of social experiments.

to those of randomized-out non-participants. The mean difference between participants and randomized-out nonparticipants is defined to be the effect of the program. Pursuit of "deep structural" parameters is abandoned. No elaborate statistical adjustments or arbitrary assumptions about functional forms of estimating equations are required to estimate the parameter of interest using randomized data. No complicated estimation strategy is required. Everyone understands means. Randomization ensures that there is no selection bias among participants, that is, there is no selection into or out of the program on the basis of the outcomes for the randomized sample.

Proponents of randomized social experiments implicitly make an important assumption: that randomization does not alter the program being studied. For certain evaluation problems and for certain behavioral models this assumption is either valid or innocuous. For other problems and models, it is not. A major conclusion of this study is that advocates of randomization have overstated their case for having avoided arbitrary assumptions. Evaluation by randomization makes implicit behavioral assumptions that in certain contexts are quite strong. Bias induced by randomization is a real possibility. And there is evidence that it is an important phenomenon.

In addition, advocates of randomization implicitly assume that certain mean differences in outcomes are invariably the objects of interest in performing an evaluation. In fact, there are many parameters of potential interest, only some of which can be cast into a mean-difference framework. Experimental methods *cannot* estimate median differences or other "quantile treatment effects" without invoking stronger assumptions than are required to recover means. The parameters of interest may not be defined by a hypothetical randomization, and randomized data may not be ideal for estimating these parameters.

Advocates of randomization are often silent on an important practical matter. Many social programs are multistage in nature. At what stage should randomization occur: at the enrollment, assignment to treatment, promotion, review of performance, or placement stage? The answer to this question reveals a contradiction in the case for randomized experiments. In order to use simple methods (that is, mean differences between participants and non-participants) to evaluate the effects of the various stages of a multistage program, it is necessary to randomize at each stage. Such multistage randomization has rarely been implemented, probably because it would drastically alter the program being evaluated.[8] But if only one randomization can be conducted, an evaluation of all stages of a multistage program entails the use of the very controversial econometric methodology sought to be avoided in the recent case for social experimentation.

[8] See, however, the evaluation of the ABC program: Ramey et al. (1976), which has multistage randomization.

The purpose of this paper is to clarify arguments for and against randomized social experiments. In order to focus the discussion, I first present a prototypical social program and consider what features of the program are of interest to policy evaluators. In the second section, I discuss the difficulties that arise in determining program features of interest. A precise statement of the evaluation problem is given. In the following section, I state the case for simple randomization; then I consider the implicit behavioral assumptions that underlie the case and the conditions under which they hold. I also discuss what can and cannot be learned from a randomized social experiment even under ideal conditions. In the fourth section I present some indirect evidence on the validity of the assumptions for the case of a recent evaluation of the Job Training Partnership Act (JTPA). I also consider some parallel studies of their validity in randomized clinical trials literature in medicine. In the fifth section I discuss the issue of choosing the appropriate stage at which one should randomize in a multistage program. In the sixth section I discuss the tension between the new and the old cases for social experimentation. The final section summarizes the argument.

12.2 Questions of Interest in Evaluating a Prototypical Social Program

The prototype considered here is a manpower training program similar to the JTPA program described by Heckman et al. (1998). That prototypical program offered a menu of training options to potential trainees. Specific job-related skills may be learned as well as general skills (such as reading, writing, arithmetic). Remedial general training may precede specific training. Job placement may be offered as a separate service independently of any skill acquisition or after completion of such an activity. Some specific skill programs entail working for an employer at a subsidized wage (that is, on-the job-training).

Individuals who receive training proceed through the following steps: they (1) apply; (2) are accepted; (3) are placed in specific training sequence; (4) are reviewed; (5) are certified in a skill; and (6) are placed with employer. For trainees receiving on-the-job training, steps (3)–(6) are combined, although trainees may be periodically reviewed during their training period. Individuals may drop out or be rejected at each stage.

Training centers were paid by the U.S. government on the basis of the quality of the placement of their trainees. Quality was measured in part by the wages received over a specified period of time after trainees complete their training program (for example, six months). Managers thus had an incentive to train persons who were likely to attain high-quality placement and who can achieve the status at low cost to the center. Trainees received compensation (subsidies) while in the program. Training centers recruited trainees through a variety of promotional schemes.

There are many questions of interest to program evaluators. The question that receives the most attention is the effect of training on the trained:

Q-1 What is the effect of training on the trained?

This is the "bottom line" stressed in many evaluations. When the costs of a program are subtracted from the answer to Q-1, and returns are appropriately discounted, the net benefit of the program is produced for a fixed group of trainees.

But there are many other questions that are also of potential interest to program evaluators, such as:

Q-2 What is the effect of training on randomly assigned trainees?

The answer to Q-2 would be of great interest if training were mandated for an entire population, as in workfare programs that force welfare recipients to take training. Other questions of interest concern application decisions:

Q-3 What is the effect of subsidies (and/or advertising, and/or local labor market conditions, and/or family income, and/or race, sex) on application decisions?

Q-4 What are the effects of center performance standards, profit rates, local labor market structure, and governmental monitoring on training center acceptance of applicant decisions and placement in specific programs?

Q-5 What are the effects of family background, center profit rates, subsidies, and local labor market conditions on the decision to drop out from a program and the length of time taken to complete the program?

Q-6 What are the effects of labor market conditions, subsidies, profit rates, and so forth on placement rates and wage and hour levels attained at placement?

Q-7 What is the cost of training a worker in the various possible ways?

Answers to all of these questions, and refinements of them, are of potential interest to policy-makers. The central evaluation problem is how to obtain convincing answers to them.

12.3 The Evaluation Problem

To characterize the essential features of the evaluation problem, it is helpful to concentrate on only on a few of the questions listed above. I focus attention on questions Q-1 and Q-2 and a combination of the ingredients in questions Q-3 and Q-4:

Q-3' What are the effects of the variables listed in Q-3 and Q-4 on application and enrollment of individuals?

To simplify the analysis, I assume throughout the discussion in this section that there is only one type of treatment administered by the program, so determining assignment to treatment is not an issue. I assume that there is no attrition from the program and that length of participation in the program is fixed. These assumptions would be true if, for example, the ideal program occurs at a single

instant in time and gives every participant the same "dose," although the response to the dose may differ across people. I also assume absence of any interdependence among units resulting from common, site-specific unobservables or feedback effects.[9]

This paper does not focus exclusively or even mainly on "structural estimation" because it is not advocated in the recent literature on social experiments and because a discussion of that topic raises additional issues that are not germane here. Structural approaches require specification of a common set of characteristics and a model of program participation and outcomes to describe all programs of potential interest. They require estimating responses to variations in characteristics that describe programs not yet put in place. This in turn requires specification and measurement of a common set of characteristics that underlie such programs.

The prototypical structural approach is well illustrated in the early work on estimating labor supply responses to negative income tax programs. Those programs operated by changing the wage level and income level of potential participants. Invoking the neoclassical theory of labor supply, if one can determine the response of labor supply to changes in wages and income levels (the "substitution" and "income" effects, respectively), one can also determine who would participate in a program (see, for example, Ashenfelter 1983). Thus from a common set of parameters, one can simulate the effect of *all possible* NIT programs on labor supply.

It is for this reason that early advocates of social experiments sought to design experiments that would give maximal sample independent variations in wage and income levels across subjects so that precise estimates of wage and income effects could be obtained. Cain and Watts (1973) argued that in cross-section data, variation in wages and income was sufficiently small that it was difficult, if not impossible, to estimate separate wage and income effects on labor supply.

The structural approach is very appealing when it is credible. It focuses on essential aspects of response to programs. But its use in practice requires invoking strong behavioral assumptions in order to place diverse programs on a common basis. In addition, it requires that the common characteristics of programs are able to be measured. Both the problems and the behavioral assumptions required in the structural approach raise issues outside the scope of this paper. I confine most of my attention to the practical—and still very difficult—problem of evaluating the effect of existing programs and the responses to changes in parameters of these programs that might affect program participation.

[9] This is Rubin's "SUTVA" assumption (see Holland 1986). It is widely invoked in the literature in econometrics and statistics, even though it is often patently false (see Heckman, Lochner and Taber 1998).

12.3.1 A Model of Program Evaluation

To be more specific, define variable $D = 1$ if a person participates in a hypothetical program; $D = 0$ otherwise. If a person participates, she/he receives outcome Y_1; otherwise she/he receives Y_0. Thus the observed outcome Y is:

$$\begin{aligned} Y &= Y_1 \text{ if } D = 1 \\ Y &= Y_0 \text{ if } D = 0 \end{aligned} \tag{12.1}$$

A crucial feature of the evaluation problem is that we do not observe the same person in both states. This is called the "problem of causal reference" by some statisticians (see, for example, Holland 1986). Let Y_1 and Y_0 be determined by X_1 and X_0 respectively. Presumably X_1 induces relevant aspects of the training received by trainees. X_0 and X_1 may contain background and local labor market variables. We write functions relating those variables to Y_0 and Y_1 respectively:

$$Y_1 = g_1(X_1), \tag{12.2a}$$
$$Y_0 = g_0(X_0), \tag{12.2b}$$

In terms of more familiar linear equations, (12.2a) and (12.2b) may be specialized to

$$Y_1 = X_1\beta_1 \tag{12.2a'}$$

and

$$Y_0 = X_0\beta_0 \tag{12.2b'}$$

respectively.
Let Z be variables determining program participation. If

$$Z \in \Psi, D = 1; \quad Z \notin \Psi, D = 0, \tag{12.3}$$

where Ψ is a set of possible Z values. If persons have characteristics that lie in set Ψ, they participate in the program; otherwise they do not. Included among the Z are characteristics of persons and their labor market opportunities as well as characteristics of the training sites selecting applicants. In order to economize on symbols, I represent the entire collection of explanatory variables by $C = (X_0, X_1, Z)$. If some variable in C does not appear in X_1 or X_0, its coefficient or associated derivative in g_1 or g_0 is set to zero for all values of the variable.

If one could observe all of the components of *C* for each person in a sample, one might still not be able to determine g_1, g_0 and Ψ. The available samples might not contain sufficient variation in the components of these vectors to trace out g_0, g_1 or to identify set Ψ. It was a "multicollinearity" problem (in income and wage variables needed to determine labor supply equations) and a lack of sample variation in income that partly motivated the original proponents of social experiments in economics.

Assuming sufficient variability in the components of the explanatory variables, one can utilize data on participants to determine g_1, on non-participants to determine g_0, and the combined sample to determine Ψ. With knowledge of these functions and sets, one can readily answer evaluation problems Q-1, Q-2, and Q-3' (provided that the support of the $X_1, X_0,$ and Z variables in the sample covers the support of these variables in the target populations of interest). It would thus be possible to construct Y_1 and Y_0 for each person and to estimate the gross gain to participation for each participant or each person in the sample. In this way questions Q-1 and Q-2 can be fully answered. From knowledge of Ψ it is possible to answer fully question Q-3' for each person.

As a practical matter, analysts do not observe all of the components of *C*. The unobserved components of these outcomes and enrollment functions are a major source of evaluation problems. It is these missing components that motivate treating Y_1, Y_0, and *D* as random variables, conditional on the available information. This intrinsic randomness rules out a strategy of determining Y_1 and Y_0 for each person. Instead, a statistical approach is adopted that focuses on estimating the joint distribution of Y_1, Y_0, D conditional on the available information or some features of it.

Let subscript *a* denote available information. Thus, C_a contains the variables *available* to the analyst thought to be legitimate for determining $Y_1, Y_0,$ and *D*. These variables may consist of some components of *C* as well as proxies for the missing components.

The joint distribution of Y_1, Y_0, D given $C_a = c_a$ is

$$F(y_0, y_1, d \mid c_a) = \Pr(Y_0 \leq y_0, Y_1 \leq y_1, D = d \mid C_a = c_a), \tag{12.4}$$

where I follow convention by denoting random variables by uppercase letters and their realization by lowercase letters. If (4) can be determined, and the distribution of C_a is known, it is possible to answer questions Q-1, Q-2, and Q-3' in the following sense: one can determine the population distribution of Y_0, Y_1 and the population *distribution* of the gross gain from the program participation,

$$\Delta = Y_1 - Y_0,$$

and one can write out the probability of the event $D = d$ given Z_a.

12.3.2 The Parameters of Interest in Program Evaluation

We can answer Q-1 if we can identify

$$F(y_0, y_1 \mid D = 1, c_a),$$

and hence

$$F(\delta \mid D = 1, c_a)$$

(the distribution of the effect of treatment on the treated, where δ is the lower-case version of Δ corresponding to realized values of Δ). One can answer Q-2 if we know

$$F(y_0, y_1 \mid c_a), \tag{12.5}$$

which can be produced from (12.4) and the distribution of the explanatory variables by elementary probability operations. In this sense, one can determine the gains from randomly moving a person from one distribution, $F(y_0 \mid c_a)$ to another $F(y_1 \mid c_a)$. The answer to Q-3' can be achieved by computing from (12.4) the probability of participation:

$$\Pr(D = 1 \mid c_a) = F(d \mid c_a).$$

In practice, comparisons of means occupy most of the attention in the literature, although medians, or other quantiles, are also of interest. Much of the literature *defines* the answer to Q-1 as

$$E(\Delta \mid D = 1, c_a) = E(Y_l - Y_0 \mid D = 1, c_a) \tag{12.6}$$

and the answer to Q-2 as

$$E(\Delta \mid c_a) = E(Y_l - Y_0 \mid c_a), \tag{12.7}$$

although in principle, knowledge of the full distribution of Δ, or some other features besides the mean (for example, the median), might be desirable.

Even if the means in (12.6) and (12.7) were zero, it is of interest to know what fraction of participants or of the population would benefit from a program. This would require knowledge of $F(\delta \mid D = 1, c_a)$ or $F(\delta \mid c_a)$, respectively. In order to ascertain the existence of "cream skimming" (the phenomenon that training sites

select the best people into a program—those with high values of Y_0 and Y_1)—it is necessary to know the correlation or stochastic dependence between Y_1 and Y_0. This would require knowledge of features of

$$F(y_1, y_0 \mid D = 1, c_a)$$

or

$$F(y_1, y_0 \mid c_a),$$

other than the means of Y_1 and Y_0. To answer many questions, knowledge of mean differences is inadequate or incomplete.

Determining the joint distribution (12.4) is a difficult problem. In the next section, I show that randomized social experiments of the sort posed in the recent literature do not produce data sufficient for this task.

The data routinely produced from social program records enable analysts to determine

$$F(y_1 \mid D = 1, c_a),$$

the distribution of outcomes for participants, and

$$F(y_0 \mid D = 0, c_a),$$

the distribution of outcomes for non-participants, and they are sometimes sufficiently rich to determine

$$\Pr(D = 1 \mid c_a) = F(d \mid c_a),$$

the probability of participation. But unless further information is available, these pieces of information do not suffice to determine (12.4). By virtue of (12.1), there are no data on both components of (y_1, Y_0) for the same person. In general, for the same values of $C_a = c_a$

$$F(y_0 \mid D = 1, c_a) \neq F(y_0 \mid D = 0, c_a) \tag{12.8a}$$

and

$$F(y_1 \mid D = 1, c_a) \neq F(y_1 \mid D = 0, c_a), \tag{12.8b}$$

which gives rise to the problem of selection bias in the outcome distributions. The more common statement of the selection problem is in terms of means:

$$E(\Delta \mid D=1, c_a) \neq E(Y_1 \mid D=1, c_a) - E(Y_0 \mid D=0, c_a) \tag{12.9a}$$

$$E(\Delta \mid c_a) \neq E(Y_1 \mid c_a) - E(Y_0 \mid c_a), \tag{12.9b}$$

that is, persons who participate in a program are different people from persons who do not participate in the sense that the mean outcomes of participants in the non-participation state would be different from those of non-participants even after adjusting for C_a.

Many methods have been proposed for solving the selection problem either for means or for entire distributions. Heckman and Honoré (1990), Heckman and Robb (1985, 1986), Heckman (1990a, 1990b), and Heckman, Smith, and Clements (1997) offer alternative comprehensive treatments of the various approaches to this problem in econometrics and statistics. Some untestable a priori assumptions must be invoked to recover the missing components of the distribution. Constructing these counterfactuals inevitably generates controversy.

LaLonde (1986) and Fraker and Maynard (1987) have argued that these controversies are of more than academic interest. In influential work analyzing randomized experimental data using non-experimental methods, these authors produce a wide array of estimates of impacts of the same program using different non-experimental methods. They claim that there is no way to choose among competing non-experimental estimators.

Heckman and Hotz (1989) reanalyze their data and demonstrate that their claims are greatly exaggerated. Neither set of authors performed standard model specification tests for their non-experimental alternative estimates. When such tests are performed, they select non-experimental models that reproduce the inference obtained by experimental methods.

There is, nonetheless, a kernel of truth in the criticism of LaLonde (1986) and Fraker and Maynard (1987). Each test of a non-experimental model proposed by Heckman and Hotz (1989) has limitations. Test of overidentifying features of a model can be rendered worthless by changing the model to a just-identified form, a criticism that also arises in application of the Durbin–Wu–Hausman test.[10]

All non-experimental methods are based on some maintained, untestable assumption. The great source of appeal of randomized experiments is that they *appear* to require no assumptions. In the next section, I demonstrate that the case for randomized evaluations rests on unstated assumptions about the problem of

[10] See Durbin (1954); Wu (1973); Hausman (1978).

interest, the number of stages in a program, and the responses of agents to randomization. These assumptions are different from but not necessarily more credible than the assumptions maintained in the non-experimental econometrics and statistics literatures.

12.4 The Case For and Against Randomized Social Experiments

The case for randomized social experiments is almost always stated within the context of obtaining answers to question Q-1 and Q-2—the "causal problem" as defined by statisticians (see Fisher 1935; Cox 1958; Rubin 1978; Holland 1986). From this vantage point, the participation equation that answers Q-3' is a "nuisance function" that may give rise to a selection problem. Simple randomization makes treatment status statistically independent of (Y_1, Y_0, C).

To state the case for randomization most clearly, it is useful to introduce a variable A indicating actual participation in a program:

$$\begin{aligned} A = &\quad 1 \text{ if a person participates} \\ = &\quad 0 \text{ otherwise} \end{aligned}$$

and separate it from variable D indicating who would have participated in a program in a non-experimental regime. Let D^* denote a variable indicating if an agent is at risk for randomization (that is, if the agent applied and was accepted in a regime of random selection):

$$\begin{aligned} D^* = &\quad 1 \text{ if a person is at risk for randomization} \\ = &\quad \text{otherwise.} \end{aligned}$$

In the standard approach, randomization is implemented at a stage when D^* is revealed. Given $D^* = 1, A$ is assumed independent of (Y_0, Y_1, C), so

$$F(y_0, y_1, c, a \mid D^* = 1) = F(y_0, y_1, c \mid D^* = 1) F(a \mid D^* = 1).$$

More elaborate randomization schemes might be implemented but are rarely proposed.

Changing the program enrollment process by randomly denying access to individuals who apply and are deemed suitable for a program may make the distribution of D^* different from D. Such randomization alters the information set of potential applicants and program administrators unless neither is informed about the possibility of randomization—an unlikely event for an ongoing program or

for one-shot programs in many countries such as the United States where full disclosure of programs operating rules is required by law. Even if it were possible to surprise potential trainees, it would not be possible to surprise training centers administering the program. (Recall that D^* is the outcome of joint decisions by potential trainees and training centers.) The conditioning set determining D^* differs from that of D by the inclusion of the probability of selection $(p = \Pr(A = 1))$, that is, it includes the effect of randomization on agent and center choices.

Proponents of randomization invoke the assumption that

$$\Pr(D = 1 \mid c) = \Pr(D^* = 1 \mid c, p), \tag{AS-1}$$

or assume that it is "practically" true.[11]

There are many reasons to suspect the validity of this assumption. If individuals who might have enrolled in a non-randomized regime make plans anticipating enrollment in training, adding uncertainty at the acceptance stage may alter their decision to apply or to undertake activities complementary to training. Risk-averse persons will tend to be eliminated from the program. Even if randomization raises agent utility,[12] behavior will be altered. If training centers must randomize after a screening process, it might be necessary for them to screen more persons in order to reach their performance goals, and this may result in lowered trainee quality. Degradation in the quality of applicants might arise even if slots in a program are rationed. Randomization may solve rationing problems in an equitable way if there is a queue for entrance into the program, but it also may alter the composition of the trainee pool.

Assumption (AS-1) is entirely natural in the context of agricultural and biological experimentation in which the Fisher model of randomized experiments was originally developed. However, the Fisher model is potentially misleading paradigm for social science. Humans act purposively, and their behavior is likely to be altered by introducing randomization in their choice environment. The Fisher model may be ideal for the study of fertilizer treatments on crop yields. Plots of ground do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated. Commercial manufacturers of fertilizer can be excluded from selecting favorable plots of ground in an agricultural experimental setting in a way that training center managers cannot be excluded from selecting favorable trainees in a social science setting.

[11] Failure of this assumption is an instance of the Marschak (1953)—Lucas (1976) Critique applied to social experimentation. It is also an instance of a "Hawthorne" effect. See Cook and Campbell (1979).

[12] This can arise even if agents are risk averse by convexifying a non-covex problem. See Arnott and Stiglitz (1988).

If (AS-1) is true,

$$F(y_1, c \mid A=1) = F(y_1, c \mid D^{\star}=1) = F(y_1, c \mid D=1), \tag{12.10a}$$

$$F(y_0, c \mid A=0) = F(y_0, c \mid D^{\star}=1) = F(y_0, c \mid D=1), \tag{12.10b}$$

$$E(Y_1 \mid A=1) - E(Y_0 \mid A=0) = E(\Delta \mid D=1). \tag{12.11}$$

Simple mean difference estimators between participants and randomized-out non-participants answer question Q-1 stated in terms of means, at least for large samples. The distribution of explanatory variables C is the same in samples conditioned on A. The samples conditioned on $A=1$ and $A=0$ are thus balanced.

In this sense, randomized data are "ideal." People untrained in statistics—such as politicians and program administrators—understand means, and no elaborate statistical adjustments or functional form assumptions about a model are imposed on the data. Moreover (12.11) *may* be true even if (AS-1) is false.

This is so for the widely used dummy endogenous variable model (Heckman 1978). For that case,

$$Y_1 = \alpha + Y_0. \tag{12.12}$$

This model is termed the "fixed treatment effect for all units model" in the statistics literature. (See Cox 1958.) That model writes

$$Y_1 = g_1(x_1) = \alpha + g_0(x_0) = \alpha + Y_0,$$

so the effect of treatment is the same for everyone. In terms of the linear regression model of (2a') and (2b'), this model can be written as $X_1\beta_1 = \alpha + X_0\beta_0$. Even if (AS-1) is false, (12.11) is true because

$$\begin{aligned}
& E(Y_1 \mid A=1) - E(Y_0 \mid A=0) \\
= \; & E(\alpha + Y_0 \mid A=1) - E(Y_0 \mid A=0) \\
= \; & \alpha + E(Y_0 \mid D^{\star}=1) - E(Y_0 \mid D^{\star}=1) \\
= \; & \alpha \\
= \; & E(\Delta \mid D=1) \\
= \; & E(\Delta).
\end{aligned}$$

The dummy endogenous variable model is widely used in applied work. Reliance on this model strengthens the popular case for randomization. Q-1 and Q-2 have the same answer in this model, and randomization provides a convincing way to answer both.

The requirement of treatment outcome homogeneity can be weakened and (12.11) can still be justified if (AS-1) is false. Suppose there is a random response model (sometimes called a random effects model):

$$Y_1 = Y_0 + (\alpha + \Xi), \tag{12.13a}$$

where Ξ is an individual's idiosyncratic response to treatment after taking out a common response α and

$$E(\Xi \mid D) = 0, \tag{12.13b}$$

then (12.11) remains true. If potential trainees and training centers do not know the trainees' gain from the program in advance of their enrollment in the program, and they use $\alpha + \Xi$ in making participation decisions, then (12.11) is still satisfied. Thus, even if responses to treatments are heterogeneous, the simple mean-difference estimator obtained from experimental data may still answer the mean-difference version of Q-1.

It is important to note how limited are the data obtained from an "ideal" social experiment (that is, one that satisfies (AS-1)). Without invoking additional assumptions, one cannot estimate the distribution of Δ conditional or unconditional on $D = 1$. One cannot estimate the median of Δ nor can one determine the empirical importance of "cream skimming" (the stochastic dependence between Y_0 and Y_1) from the data, unless one makes the extreme assumption of rank invariance (i.e., that the rank of persons in the Y_1 distribution are the same in the Y_0 distribution).[13] Both experimental and non-experimental data are still plagued by the fundamental problem that one cannot observe Y_0 and Y_1 for the same person. Randomized experimental data of the type proposed in the literature only facilitate simple estimation of one parameter,

$$E(\Delta \mid D = 1, c).$$

Assumptions must be imposed to produce additional parameters of interest even from ideal experimental data. Answers to most of the questions listed in the first section still require application of econometric procedures with their attendant controversial assumptions.

If assumption (AS-1) is not satisfied, the final equalities in (12.10a) and (12.10b) are not satisfied, and in general

$$E(Y_1 \mid A = 1) - E(Y_0 \mid A = 0) \neq E(\Delta \mid D = 1).$$

[13] See Heckman, Smith, and Clements (1997).

Moreover, the data produced by the experiment will not enable analysts to assess the determinants of participation in a non-randomized regime because the application and enrollment decision processes will have been altered by randomization; that is,

$$\Pr(D=1\,|\,c) \neq \Pr(D^{\star}=1\,|\,c,p),$$

unless $p=1$. Thus, experimentation will not produce data to answer question Q-3′ unless randomization is a permanent feature of the program being evaluated.

In the general case in which agents' response to programs is heterogenous ($\Xi \neq 0$) and agents anticipate this heterogeneity (more precisely, Ξ is not stochastically independent of D), assumption (AS-1) plays a crucial role in justifying randomized social experiments. While (AS-1) is entirely non-controversial in some areas of science—such as in agricultural experimentation where the original Fisher model was developed—it is more problematic in social settings. It may produce clear answers to the wrong question and may produce data that cannot be used to answer crucial evaluation questions, even when question Q-1 can be clearly answered.

12.5 Evidence on Randomization Bias

Violations of assumption (AS-1) in general make the evidence from randomized social experiments unreliable. How important is this theoretical possibility in practice? Surprisingly, very little is known about the answer to this question for the social experiments conducted in economics. This is so because, except for one program, randomized social experimentation has only been implemented on "pilot projects" or "demonstration projects" designed to evaluate new programs without precedent. The possibility of disruption by randomization cannot be confirmed or denied on data from these experiments. In one program evaluated by randomization, participation was compulsory for the target population (Doolittle and Traeger 1990). Hence, randomization did not affect applicant pools or assessments of applicant eligibility by program administrators.

Fortunately, there is some information on this question, although it is indirect. In response to the wide variability in estimates of the impact of manpower programs derived from non-experimental estimators by LaLonde (1986) and Fraker and Maynard (1987), the U.S. Department of Labor financed a large-scale experimental evaluation of the large-scale Job Training Partnership Act (JTPA), which was the main vehicle for providing government training in the United

States. Randomized evaluation was implemented in a variety of sites. The organization implementing this experiment—the Manpower Demonstration Research Corporation (MDRC)—is an ardent and effective advocate for the use of randomization as a method for evaluating social programs.

A report by MDRC (Doolittle and Traeger 1990) gives some information from which it is possible to do a rough revealed preference analysis.[14] Job training in the United States in the late 1980s and early 1990s was organized through geographically decentralized centers. These centers received incentive payments for placing unemployed persons and persons on welfare in "high-paying" jobs. The participation of centers in the experiment was not compulsory. Funds were set aside to compensate job centers for the administrative costs of participating in the experiment. The funds set aside range from 5 percent to 10 percent of the total operating costs of the centers.

In attempting to enroll geographically dispersed sites, MDRC experienced a training center refusal rate in excess of 90 percent. The reasons for refusal to participate are given in Table 12.1. (The reasons stated there are not mutually exclusive.) Leading the list are ethical and public relation objections to randomization. Major fears (items 2 and 3) were expressed about the effects of randomization on the quality of applicant pool, which would impede the profitability of the training centers. By randomizing, the centers had to widen the available pool of persons deemed eligible, and there was great concern about the effects of this widening on applicant quality—precisely the behavior ruled out by assumption (AS-1). In attempting to entice centers to participate, MDRC had to reduce the randomized rejection probability from $\frac{1}{2}$ to as low as $\frac{1}{6}$ for certain centers. The resulting reduction in the size of the control sample impairs the power of statistical tests designed to test the null hypothesis of no program effect. Compensation was expanded sevenfold in order to get any centers to participate in the experiment. The MDRC analysts concluded:

> Implementing a complex random assignment research design in an ongoing program providing a variety of services does inevitably change its operation in some ways . . . The most likely difference arising from a random assignment field study of program impacts . . . is a change in the mix of clients served. Expanded recruitment efforts, needed to generate the control group, draw in additional applicants who are not identical to the people previously served. A second likely change is that the treatment categories may somewhat restrict program staff's flexibility to change service recommendations.
>
> (Doolittle and Traeger 1990: 121)

[14] Hotz (1992) also summarizes their discussion.

Table 12.1 Percentage of local JTPA agencies citing specific concerns about participating in the experiment

Concern	Percentage of training centers citing the concern
1. Ethical and public relations implications of:	
a. Random assignment in social programs	61.8
b. Denial of services to controls	54.5
2. Potential negative effect of creation of a control group on achievement of client recruitment goals	47.8
3. Potential negative impact on performance standards	25.4
4. Implementation of the study when service providers do intake	21.1
5. Objections of service providers to the study	17.5
6. Potential staff administrative burden	16.2
7. Possible lack of support by elected officials	15.8
8. Legality of random assignment and possible grievances	14.5
9. Procedures for providing controls with referrals to other services	14.0
10. Special recruitment problems for out-of-school youth	10.5
Sample size	228

Notes: Concerns noted by fewer than 5 percent of the training centers are not listed. Percentages may add to more than 100.0 because training centers could raise more than one concern.

Source: Based on responses of 228 local JTPA agencies contacted about possible participation in the National JTPA Study. From Doolittle and Traeger (1990). Copyright 1989, 1990 by the Manpower Demonstration Research Corporation and used with its permission.

These authors go on to note that "some [training centers], because of severe recruitment problems or up-front services, cannot implement the type of random assignment model needed to answer the various impact questions without major changes in procedures" (p. 123).

During the experiment conducted at Corpus Christi, Texas, center administrators successfully petitioned the government of Texas for a waiver of its performance standards on the ground that the experiment disrupted center operations. Self-selection likely guarantees that participant sites are the least likely sites to suffer disruption. Such selective participation in the experiment calls into question the validity of the experimental estimates as a statement about the JTPA system as a whole. At least the data can be used to provide a lower-bound estimate of the major impact of disruption.

Randomization is also controversial in clinical trials in medicine which are sometimes held up as a paragon for empirical social science.[15] The ethical problem

[15] See, for example, Ashenfelter and Card (1985).

raised by the manpower training centers of denying equally qualified persons access to training has its counterpart in the application of randomized clinical trials. For example, Joseph Palca, writing in *Science* (1989), notes that AIDS patients denied potentially life-saving drugs took steps to undo random assignment. Patients had the pills they were taking tested to see if they were getting a placebo or an unsatisfactory treatment, and were likely to drop out of the experiment in either case or to seek more effective medication, or both. In the MDRC experiment, in some sites qualified trainees found alternative avenues for securing exactly the same training presented by the same subcontractors by using other methods of financial support.

Writing in the *Journal of the American Medical Association*, Kramer and Shapiro (1984: 2739) note that subjects in drug trials were less likely to participate in randomized trials than in non-experimental studies. They discuss one study of drugs administered to children afflicted with a disease. The study had two components. The non-experimental phase of the study had a 4 percent refusal rate, while 34 percent of a subsample of the same parents refused to participate in a randomized subtrial, although the treatments were equally non-threatening.

These authors cite evidence suggesting that non-response to randomization is selective. In a study of treatment of adults of cirrhosis, no effect of the treatment was found for participants in a randomized trial. But the death rates for those randomized out of the treatment were substantially lower than among those individuals who refused to participate in the experiment, despite the fact that both groups were administered the same alternative treatment.

This evidence qualifies the case for randomized social experimentation. Where feasible, it may alter the program being studied. For many social programs it is not a feasible tool for evaluation.

12.6 At What Stage Should Randomization Be Implemented?

Thus far, I have deliberately abstracted from the multistage feature of most social programs. In this section, I briefly consider the issue of the choice of the stage in a multistage program at which randomization should be implemented.

In principle, randomization could be performed to evaluate outcomes at each stage. The fact that multiple randomization has rarely been performed likely indicates that it would exacerbate the problem of randomization bias discussed in the two previous sections. Assuming the absence of randomization bias, if only one randomization is to be performed, at what stage should it be placed? One obvious answer is at the stage where it is least disruptive, although that stage is not so easy to determine in the absence of considerable information about the process being

studied. If randomization is performed at one stage, non-experimental "econometric" or "statistical" estimators are required to evaluate outcomes attributable to participation at all other stages. This accounts for the sometimes very complicated (Ham and LaLonde 1990) or controversial (Cain and Wissoker 1990; Hannan and Tuma Brandon 1990) analyses of randomized experimental data that have appeared in the recent literature.

Moreover, for some of the questions posed at the beginning of the paper, it is not obvious that randomization is the method of choice for securing convincing answers. Many of the questions listed there concern the response of trainees and training centers to variations in constraints. While enhanced variation in explanatory variables (in a sense, made precise by Conlisk 1973) facilitates estimation of response functions, there is no reason why randomized allocations are desirable or optimal for this purpose.

Thus, if we seek to enhance our knowledge of how family income determines program participation, it is not obvious that randomly allocated allotments of family income supplements are a cost-effective or optimal substitute for non-experimental optimal sample design strategies that oversample family income at the extremes of the eligible population.[16]

If we seek to enhance our knowledge about how local labor market conditions affect enrollment, retention, and training-center acceptance and placement decisions, variation across training sites and these conditions would be desirable. It is not obvious that randomization is the best way to secure this variation.

Randomization in eligibility for the program has been proposed as an alternative to randomization at enrollment. This is sometimes deemed to be a more acceptable randomization point because it avoids the application and screening costs that are incurred when accepted individuals are randomized out of a program. Since the randomization is performed outside of the training center, it prevents the center from bearing the political cost of denying eligible persons the right to participate in the program. For this reason, it is thought to be less disruptive than randomization performed at some other stage.

If eligibility is randomly assigned in the population, it still encounters the problem that people self-select. Assuming that eligibility does not disrupt the fundamental program parameters, the simple mean-difference parameter comparing the eligible with the ineligible identifies $E(Y_1 - Y_0 \mid D = 1)P$, where P is the probability of participation in the program through voluntary selection in the program in the absence of an experiment. Dividing by P, one can identify treatment on the treated.

[16] This remark assumes a linear model. For optimal designs in nonlinear models see, for example, Silvey (1980).

12.7 The Tension between the Case for Social Experiments as a Substitute for Behavioral Models and Social Experiments as Supplementary Source of Information

There is an intellectual tension between the optimal experimental design point of view and the simple mean difference point of view toward social experiments. The older optimal experimental design point of view stresses explicit models and the use of experiments to recover parameters of behavioral or "structural" models. The simple randomization point of view seeks to bypass models and produces—under certain conditions—a clean answer to one question (Q-1): does the program work for participants? The two points of view can be reconciled if one is agnostic about the prior information at the disposal of analysts to design experiments (see Savage 1962). However, the benefits of randomization are less apparent when the goal is to recover trainee participation and continuation functions than if it is to recover the distribution of program outcome measures.

The potential conflict between the objectives of experimentation as a means of obtaining better estimates of a behavioral model and experimentation as a method for producing simple estimators of mean program impacts comes out forcefully when we consider using data from randomized experiments to estimate a behavioral model. To focus on main points, consider a program with two stages. $D_1 = 1$ if a person completes stage one; $= 0$ otherwise. $D_2 = 1$ if a person completes stage two; $= 0$ otherwise. Suppose that outcome *Y* can be written in the following form:

$$Y = \theta_0 + \theta_1 D_1 + \theta_2 D_1 D_2 + U. \tag{12.14}$$

The statistical problem is that D_1 and D_2 are stochastically dependent on *U*. Randomizing at stage one makes D_1 independent of *U*. It does not guarantee that $D_1 D_2$ is stochastically independent of *U* because participation in stage 2 is contingent on participation at stage 1.

The simple mean-difference estimator, comparing outcomes of stage one completers with outcomes of those randomized out, estimates, in large samples,

$$E(Y \mid D_1 = 1) - E(Y \mid D_1 = 0) = \theta_1 + \theta_2 E(D_2 \mid D_1 = 1).$$

In order to estimate θ_2 or θ_1 to estimate marginal effects of program completion at each stage, it is necessary to find an instrumental variable for $D_1 D_2$.

Randomization on one coordinate only eliminates the need for one instrument to achieve this task. The appropriate stage at which the randomization should be implemented is an open question. The trade-off between randomization as an instrumental variable and better non-experimental sample design remains to be investigated. The optimal design of an experiment to estimate the parameters of

(12.14) in general would not entail simple randomization at one stage. The data generated as a by-product of a one-shot randomization are only ideal for the estimation of models like (12.14) in the limited sense of requiring one less instrumental variable to consistently estimate θ_1 or θ_2, although this is a real benefit.

12.8 Summary of the 1992 Paper

This paper critically examines the case made in the First Awakening for randomized social experimentation as a method for evaluating social programs. The method produces convincing answers to certain policy questions under strong assumptions about the behavior of agents and the questions of interest to program evaluators.

The method is ideal for evaluating social programs if attention focuses on estimating the *mean* effect of treatment on outcomes of the treated and if one of the following set of assumptions holds:

(AS-1) There is no effect of randomization on participation decisions;

or

(AS-2) If there is an effect of randomization on participation decisions, either

(a) the effect of treatment is the same for all participants or
(b) if agents differ in their response to treatments, their idiosyncratic responses to treatment do not influence their participation decisions.

If attention focuses on other features of social programs such as the determinants of participation, rejection, or continuation decisions, randomized data possesses no comparative advantage over stratified, non-randomized data. Even if (AS-1) is true, experimental data cannot be used to investigate the distribution of program outcomes or their median without invoking additional "statistical" or "econometric" assumptions. In a multistage program, randomized experimental data produce a "clean" (mean-difference) estimator of program impact only for outcomes defined conditionally on the stage(s) where randomization is implemented. Statistical methods with their accompanying assumptions must still be used to evaluate outcomes at other stages and marginal outcomes for each stage.

Under assumptions that ensure that it produces valid answers, the randomized experimental method bypasses the need to specify elaborate behavioral models. However, this makes experimental evidence an inflexible vehicle for predicting outcomes in environments different from those used to conduct the experiment. Interpolation and extrapolation replace model-based forecasting. However, such curve-fitting procedures may produce more convincing forecasts than ones produced from a controversial behavioral model.

Assumption (AS-1) is not controversial in the context of randomized agricultural experimentation. This was the setting in which the Fisher (1935) model of experiments was developed. That model is the intellectual foundation for recent case for social experiments, although the recent literature in economics often misattributes it to statisticians of the 1970s. Assumption (AS-1) is more controversial even in the context of randomized clinical trials in medicine. Human agents may respond to randomization, and these responses potentially threaten the reliability of experimental evidence. The evidence on randomization bias presented earlier calls into question the validity of (AS-1).

If that assumption is not valid, and if the program participants respond differently to common treatments, and those differences at least partly determine program participation decisions (so that (AS-2) is false), experimental methods do not even estimate the mean effect of treatment on the treated. In this case, randomized experimental methods answer the wrong question unless randomization is a permanent feature of the social program being evaluated. Data from randomized experiments cannot be used to estimate program participation, enrollment, and continuation equations for ongoing programs.

12.9 Postscript, 2019

I stand by my discussion of the conceptual issues raised in this paper and my companion paper with Heckman and Smith (1995).[17] The points made are all valid today and have largely been ignored in the recent "Second Awakening" revival in development economics. There are many papers written after these papers that establish or reiterate the points made here. In addition to failing to learn from the past, the Randomistas are ungenerous to the true pioneers of field experiments.

In subsequent work, Heckman and Smith (1998) develop the point that self-selection into a program generates information about agent *ex ante* perceptions of program benefits.[18] These subjective evaluations are arguably more important than the "objective" evaluations (δ) emphasized by statisticians who treat "noncompliance" as a problem rather than a source of information. This information would be suppressed if persons were forced to go into treatment or control status. This point is yet one more example of the benefits of using economics to design and evaluate social programs.

Later work, Heckman et al. (2000), considers *substitution bias* as a major threat to straightforward interpretation of experiments. If agents have access to alternative programs, persons eligible to participate in a program and persons ineligible

[17] I have since amplified these points in Heckman, Ichimura, and Todd (1997), Heckman, LaLonde, and Smith (1999), and Heckman and Vytlacil (2007).

[18] Thus, as noted by Heckman and Smith (1998), the pain and suffering of a medical trial may outweigh its benefits for survival.

may choose to participate in an alternative program. The "transparent" mean difference between treatments and controls does not compare the effect of treatment with no treatment, but instead, the effect of treatment vs the best alternative, which may in fact be better than the program being evaluated. Our (2000) paper documents the pervasiveness of the problem. Kline and Walters (2016) give a recent demonstration of the problem of substitution bias. The "transparent" mean difference estimator from a recent experimental evaluation of Head Start suggested that the program had no impact on disadvantaged children. A more careful analysis accounting for substitution bias using microeconometric methods shows a strong effect. Their paper mirrors the 1978 finding of Cogan regarding the NIT.

Banerjee and Duflo (2009) respond to the points raised in my 1992 paper, as do Athey and Imbens (2017). They claim that its criticisms no longer apply due to improved survey design and implementation methodology. However, they do not discuss many basic interpretive or conceptual points made in my 1992 paper or its 1995 companion, or the inability of experimental mean difference comparisons to answer the range of policy-relevant treatment effects discussed in my papers and in subsequent research (Heckman 2008).

The literature after my 1992 paper has produced considerable evidence on the inadequacy of experimental evidence in many fields. Sanson-Fisher et al. (2007) show that experiments are fundamentally too limited in scope to consider impact evaluations, such as women's empowerment. Concato and Horwitz (2018) survey the consensus in medicine.[19] It has switched away from reliance on RCTs as the "gold standard," which they say was the party line in the 1990s in medicine. They present many papers discussing limitations of randomized experiments in medicine (Horwitz 1996; Feinstein and Horwitz 1997; Concato and Horwitz 2004; Concato 2012; Concato 2013; Horwitz and Singer 2017; Shahar 1997; Sehon and Stanley 2003; Chakravarty and Fries 2006; Worrall 2007; Rawlins 2008; Borgerson 2009; Frieden 2017). Czibor, Jimenez-Gomez, and List (2019) is a recent cautionary paper for experimental economists that reiterates the points of my 1992 paper. It highlights serious problems in experimental economics and what devout experimentalists need to be wary of.

The causal models advocated in the recent program evaluation literature are motivated by the experiment as an ideal. They do not clearly specify the theoretical mechanisms determining the sets of possible counterfactual outcomes, how hypothetical counterfactuals are realized or how hypothetical interventions are implemented except to compare "randomized" with "non-randomized" interventions. They focus on outcomes, leaving the model for selecting outcomes and the preferences of agents over expected outcomes unspecified.[20]

Those who ignore intellectual history are condemned to repeat past mistakes. The Second Wave will pass as economists relearn the lessons of the past.

[19] The "paragon" cited by Ashenfelter and Card (1985).

[20] See Heckman (2008).

Interviews

Interview with Jean-Paul Moatti and Rémy Rioux

Rémy Rioux, you head an institution (AFD) with a mission to fund development projects, programmes and policies. What does evaluation entail in your institution and what role does it play in your and your partners' work?

To clearly understand the role ascribed to evaluation, let's first situate it in the Agence Française de Développement's work cycle in the light of what I refer to in my book *Reconciliations* as the four-pillar LACE approach: listening, appraising, committing, and evaluating. The starting point for AFD's identification of a project or public policy is always the expression of a need by a partner in a country of the Global South, such as a ministry, a local authority or a cross-border cooperation body. If this need is in tune with the development aid priorities set for our Group by the French government, then we enter into a phase of listening and dialogue where AFD draws on internal and external expertise to analyze its advisability as well as its technical, financial, and institutional feasibility. This preliminary design phase is key, since it conditions the scope of the project evaluation by defining, together with the stakeholders, the project's transformative goals.

The project is then screened by a "sustainable development opinion," issued by a structure independent of our operations, which uses a composite analytic grid to estimate and rate the project's potential impacts on six aspects of sustainable development. So the evaluation focuses not just on direct material outputs, but also—and most importantly—on the accomplishment of the project's purpose. Indeed, our development actions should be assessed in relation to what they add to the local economic, environmental, and social fabric. For example, if a project aims to raise the school enrolment rate for children and improve their educational achievements, then building functional establishments that last is an important condition, but it is not enough to make the project a success. The project will only be judged successful if school attendance is high and an improvement is duly observed in pupils' skills at the end of the syllabus. Therefore, evaluation is only possible if the transformative principle has been clarified right from the design of the project and if success criteria—in the shape of methodically reported quantitative and qualitative indicators—have been explicitly identified to start with.

Lastly, a monitoring mechanism needs to be set up before project implementation gets underway to check on both the smooth progress of the funded activities and the achievement of the mid-term outcomes. In our education project, for example, this means ensuring well before the end of the project that teachers have been trained and posted to the establishments built by the project, and that inequalities of access between girls and boys are gradually narrowing. If not, corrective measures will need to be taken before project completion to meet the goals set.

In other words, evaluation plays an eminently operational role as a key component of the development practitioner's job to understand the parameters that have enabled, accelerated or put a brake on the project's progress and achievement of its objectives, draw lessons, and improve effectiveness. But evaluation is also playing an increasingly strategic role. At a time of growing needs for investment to reduce inequalities and preserve our planet's balances, the main question a donor such as AFD asks is, "How can evaluation help us to co-construct development projects with our partners and support more efficient and effective public policies for populations?" One lesson I've drawn from our experience at AFD is that evaluating our action builds on an accountability and continuous improvement approach. That's why evaluation plays a central and growing role in our institution.

Admittedly, on account of its history, the French aid evaluation long took a back seat compared with other countries and multilateral donors. Comparative studies by Laporte (2015) show that evaluation practices emerged at the same time in France and the Anglo-Saxon countries in the 1960s, with a specific approach here marked by the role of the French statistics institute (INSEE) and French cooperation agency agricultural and rural development experts. However, France fell behind in the 1980s and 1990s at a time when Anglo-Saxon donors were systematizing evaluation as their aid budgets rose.

France is making up this lost ground driven by a combination of our rising investments in sustainable development and associated needs for results. AFD's resources are gradually growing to meet the goal set by the French President to channel 0.55% of GDP into official development assistance by 2022. This budget increase naturally comes with an imperative: to be ever more accountable to government, members of parliament, and the public at large for the effectiveness of our actions. Evaluation plays an instrumental role in meeting this accountability requirement.

Among its evaluations, AFD conducts impact studies, which are geared towards a more scientific approach. How useful is this evaluative research and what do you think of RCTs (randomized control trials), which are presented today as the most rigorous approach in the field?

In my first answer, I discussed evaluations in general. Impact studies are a particular type of evaluation designed to identify, measure, and understand effects

strictly attributable to an intervention in a scientifically rigorous manner. They are based on a counterfactual approach which compares and contrasts the evolution of a beneficiary population against that of a non-beneficiary population, making sure that these two populations are indeed comparable and that the evaluated intervention is the only criterion that differentiates them.

Impact evaluations are essentially used as an ad-hoc tool to prove rather than to improve. They can serve to establish that, in general, certain types of interventions work in given settings, but impact evaluations are too heavy, long-winded, and expensive to be used in a systematic manner as an accountability and learning instrument. When it comes to accountability and learning, the development community can use lighter, harmonized approaches—which it has regularly improved over the past three decades, which we are able to use with greater agility, and which inform our dialogue with our partners based on existing local expertise.

AFD has been conducting impact studies since the early 2000s to contribute to the body of general knowledge on development. The detailed benchmark we have put together on this subject (Pamies-Sumner, 2015) shows that this effort is unique to AFD. With the exception of DFID, most bilateral funders are lagging behind in the utilization of this methodology compared with the large multilateral funders that have produced hundreds of impact studies.

AFD has also been actively contributing to the development community's methodological thinking on impact evaluations for the last fifteen years. Among the methods that can be used to improve comparability between beneficiary population and counterfactual to deduce an intervention's impact, the RCT method, advocated by Nobel laureate Esther Duflo, has seen a massive wave of popularity as a method to understand the effectiveness of development policy interventions, which this book proposes to assess. AFD played a key role in sparking this trend by supporting, back in 2005, two vast RCTs on intervention sectors that were aid headliners at the time: microcredit and health insurance. These studies enabled us to contribute to the state of knowledge on both these intervention sectors as well as on experimental methods, on which we subsequently gave a mixed review (Bernard, Delarue, and Naudet 2012). With these exercises, we contributed to the establishment of a consensus on the need to move towards multidisciplinary approaches combining quantitative and qualitative methods.

Today, the need for effective action is greater than ever. There are only ten years left to reach the Sustainable Development Goals. Public action needs to be increasingly evidence-based to respond to citizens' legitimately growing demands in a world riddled with social, political, economic, and, of course, environmental fractures. As a sustainable development platform, AFD has a duty to be practical and act efficiently, and is determined to conduct more impact evaluations. We have the means, objectively and scientifically speaking, to know what works and what doesn't work.

There remains room for improvement. As I argue in *Reconciliations*, current approaches to evaluating the impact of development aid leave a lot to be desired. Outcomes are generally measured by a necessary and useful—yet inadequate—snapshot of quantitative variables such as the number of people connected to the electricity grid, how many young girls attend school and the quantity of tons of CO2 saved. Quantitative approaches tend to be disappointing because they fail to identify the way in which a successful development project has contributed to qualitative dimensions of Agenda 2030. In this regard, defining a common framework to establish which investments are aligned with long-term sustainable trajectories and which are not is essential to improve qualitative approaches to evaluation so as to meet the objectives of the Paris Agreement and the SDGs.

As methodologies have diversified, what is important now is to use the most relevant ones to answer the questions on the subjects studied, without any theoretical preconceptions. This is why AFD teams constantly seek to combine impact studies with lighter-weight operational evaluations, developing a diversified range of measurement tools to best capture aid effectiveness. In particular, we plan on making more use of science for evaluation, working with partners such as the French National Research Institute for Sustainable Development (IRD) and the world of research in general, including by promoting research conducted by partners in the Global South.

Jean-Paul Moatti, until recently, you have been CEO of IRD. But first, as a health economist, you are well acquainted with medical trials. How do you view RCTs based on your own experience?

I have spent most of the forty years of my academic career working closely with epidemiologists and biostatisticians who long upheld randomized trials as a key tool for rationalizing medical practice on the basis of scientific evidence, but ultimately expressed concerns about relying too much on randomized design.

RCTs have long been presumed to be the ideal source for data on the effects of medical treatment. Obviously, other study designs like cohort and case control studies are used where randomization is not possible for ethical or practical reasons, as often seen with studies on environmental risk factors. However, recent years have seen growing interest in other methods for obtaining evidence for decisive action in epidemiology and public health.

Well-designed RCTs can claim to provide strong internal validity in that they evenly distribute known and unknown factors between control and intervention groups, thereby limiting the potential for confounding factors in the identification of a causal mechanism. Yet public health experts have long recognized the limitations of RCTs when it comes to their external validity and their suitability for decision-making. Now, there can be a number of reasons why an RCT lacks external validity. Generalization of findings outside the study population might be invalid. RCTs do not usually have long enough study periods or large enough

populations to be able to assess how long a treatment effect might last, as in the case of the impact of vaccines on a population's long-term immunity, or to identify rare but serious adverse treatment effects, which often become evident in the post-marketing monitoring and long-term follow-up stages. The increasingly high cost and time constraints of RCTs can prompt reliance on surrogate markers that may not correlate well with the real outcome of interest. To both restrict sample size and secure sufficient statistical power, RCTs often concentrate on high-risk groups, thereby reducing the likelihood of their relevance to broader target populations. And then most RCTs take years to plan, implement, and analyze, limiting their ability to keep pace with biomedical innovations and forcing decisions to be made about new drugs and medical devices before clinical evaluation has been completed. RCTs' time constraints also rule out their effective use for public health decisions in the event of outbreaks of epidemics. Moreover, contradictory results have often been produced by different RCTs studying the same issue, especially regarding key questions on the effectiveness of medical practice. This has led to the development of the so-called *meta-analysis* methods that combine study results and draw evidence-based conclusions. However, these methods themselves raise complex statistical problems.

On the whole, public health experts now tend to consider that today's evidence-grading systems are biased toward RCTs and may potentially sideline non-RCT data.

More recently, growing numbers of biostatisticians have recognized that randomization is not in itself an absolute guarantee of internal validity. For example, Cook (2018) lists 26 assumptions that could bias RCT results despite randomization, 22 of which are internal validity concerns: a pre-intervention group selection difference that could be mistaken for a treatment effect; the possibility that trial attrition might have differed between groups, making results highly sensitive to the treatment of missing values; bias in the choice of control group (e.g. when innovation is compared with a non-optimal current standard of care); behavioral changes induced by trial participation (e.g. in some double-blind trials with a placebo control arm, HIV-infected patients started to share their pills in order to guarantee "at least" some effective drugs to all participants), and so on. Most of these concerns revolve around one key issue: if randomization, voluntarily or involuntarily, ex-ante or ex-post (in the shape of the analysis of the trial itself), ignores prior information from theory and covariates, then it is wasteful and even unethical because it unnecessarily exposes people to possible harm in a risky experiment.

Alterations to the randomized trial's basic structure have been developed to minimize this risk with measures such as stratification, adaptive allocation, and pre-randomization to prevent imbalance in known or identified prognosis factors. Single case designs (SCDs) are used when a dependent variable of interest can be measured repeatedly over time between two points (at baseline and during

or after the intervention). Rather than using randomization of large numbers of participants, SCD researchers use careful and prescribed ordering of experimental conditions to improve internal validity by ruling out alternative explanations for treatment effects. All these attempts to improve internal validity de facto recognize that randomization does not, in practice, present any intrinsic statistical superiority for causal inference.

Donning your other hat, as an econometrician, you are presumably familiar with the methodological issues surrounding causal inference and with alternative designs to randomization to identify the causal factors of the outcome phenomenon of interest.

With these developments and the current move toward "pragmatic trials," public health experts have rediscovered an "old" econometric Bayesian argument dating back to Fisher (1926) and Savage (1962), which challenges the belief that average treatment effects estimated from RCTs are likely to be closer to the truth than those estimated in other ways. The outcome of interest of any RCT is the difference in means between the intervention and control groups, which combines the average treatment effect among those treated with the error term reflecting the randomly generated imbalance in the net effects of the other causes. RCTs provide the basis for the calculation of the size of the error but, as mentioned before, that is conditional on the caveat that no post-randomization correlation with covariates has occurred. For example, statistical significance, in RCTs as in other designs, will be threatened if there is an asymmetric distribution of individual treatment effects in the study population. Summarizing a vast body of econometric literature on the subject, Nobel Laureate in Economics Angus Deaton (Deaton and Cartwright 2018) legitimately argues that any special status for RCTs is unwarranted and concludes, "Which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known."

Econometricians are familiar with other methods designed to yield causal inference such as propensity score matching, instrumental variables, econometric modeling, and causal Bayesian networks. Obviously, all methods have to generate control groups for appropriate comparison with the intervention, but the choice of study design should remain pragmatic and depend on the issues at stake.

Another source of scepticism about randomized experiments among econometricians, dating back to another Nobel Laureate in Economics, Heckman (Heckman and Smith 1995), is precisely that information on average treatment effects may not be very useful to inform policy, since they ignore variations across intervention beneficiaries. Mean impacts may be the main point of interest in an evaluation comparing two drugs or two very simple interventions, but when it comes to multi-component policies, garnering useful lessons from an experiment implies rather identifying the reasons why some work better than others. In such cases, even a successful RCT

cannot guarantee that the established causal relation will hold in other settings or in general. Claiming that microfinance should be at the core of poverty eradication efforts or that conditional cash transfers should be the priority of health and education policies on the basis of a limited number of randomized experiments in these fields is clearly misleading and may distract from major policies designed to reduce inequalities or provide health and education for all. Interestingly enough, some proponents of randomized experiments, such as Banerjee (2015), come to a similar conclusion when they recognize, for example, that "The microcredit enthusiasts may have [. . .] overestimated the potential of businesses for the poor, both as a source of revenue and as a means of empowerment for their female owners."

Rémy Rioux, have you observed any operational returns on the evaluations for the projects you support on the ground and what developments have you seen or would you like to see?

In general, we have observed that the culture of evaluation fosters a culture of innovation. At AFD, impact assessments have generated interesting returns in that they have consolidated the culture of evaluation in our Group and among our partners in the Global South.

In Mauritania, for example, an impact evaluation on a social protection mechanism covering 40 percent of women showed that while this mechanism significantly raised the use of care and reduced inequalities, it did not reach the poor and had no significant impact on mother and child health due to the deterioration in the quality of care in the establishments (Philibert et al. 2017). As a result, the next phase of the project underwent a complete change of paradigm, taking across-the-board action on the different quality components (human resources, blood, medicine, and supervision) and operationalizing a free mechanism for the poor.

Evaluations have also fostered methodological innovations that have improved project monitoring. For instance, we now offer to help project managers and partners to use existing data right from the project appraisal stage to better estimate living conditions and access to services, and analyze household expenditure, and so on, or to use satellite images to monitor productivity, deforestation, urban development, and so forth. Ongoing project monitoring is also helped by digital monitoring methods such as Geopoppy, developed by the French National Institute for Agricultural Research (INRA), which we have used to monitor agriculture in Côte d'Ivoire and which we will even be developing as a capacity-building tool in Benin. In the same vein, AFD works with the Center for Research and Interdisciplinarity (CRI) in Haiti and Niger—which was founded by Francois Taddei and Ariel Lindner—to experiment and spread new ways of conducting research and mobilizing collective intelligence in life, learning, and digital sciences.

All these examples show that, when we can capture the impacts of our projects, we learn to share our experience, further collaborate with our beneficiaries, and

ultimately innovate with them and for them. By evaluating our impacts, we can innovate and show in concrete terms the return on investment that development policy generates. An inclusive, sustainable investment.

Jean-Paul Moatti, you too are engaged in informing policies. In addition to your academic credentials, you are a member of the UN expert panel in charge of the Global Sustainable Development Report. As such, you contributed to the first four-year global assessment report on the Sustainable Development Goals adopted by all UN member states in September 2015 and setting the 2030 international agenda for development. We are seeing an emerging concept of "sustainability science." Do you support this concept and how randomized experiments may, or may not, contribute to effective research for sustainable development?

As you know, the 17 Sustainable Development Goals (SDGs) setting the 2030 international agenda for development were adopted by all UN member states in September 2015. Although these goals remain the products of many compromises between governments and conflicting interests, this ambitious, transformative agenda benefited enormously from the emergence of what is now called "sustainability science." The US National Academy of Sciences, which started promoting sustainability science back in 2000, defines it as, "An emerging field of research dealing with the interactions between natural and social systems, and with how those interactions affect the challenge of sustainability: meeting the needs of present and future generations while substantially reducing poverty and conserving the planet's life support systems." Because it is problem-driven, this new scientific approach is by essence interdisciplinary and open to co-construction of research programmes with affected communities. It focuses on identifying the complex causal chains that generate the major environmental and social concerns threatening the Earth's future and on proposing solutions to reduce the risk of inconsistency in the implementation of the SDGs while maximizing positive synergies between them: how can agricultural productivity be increased to guarantee food security for a growing world population while reducing chemical inputs in order to limit the environmental impact and waste of resources? How can sustainable growth be promoted for absolute poverty eradication without increasing intra-country inequalities? The issues at stake are many.

Development economics should play a leading role in this sustainability science, because it uses skills and knowledge key to the translation into effective policies of evidence-based facts and experience from a large body of disciplines, from natural to social sciences, and to their adaptation to heterogeneous social, environmental and economic contexts. Randomized experiments are often not suited to this crucial interdisciplinary field, since the extrapolation and generalization of their results calls for a whole host of other information that has to come from other sources. Overestimating their role, while ignoring the limitations of

randomized design and overselling them to decision-makers, could jeopardize the contribution of economics to the conversion of current development models to sustainability. However, it would also be a mistake to underestimate the fact that randomized studies can be extremely useful, where appropriate, to determine the best practices for the SDGs among alternative intervention modalities and to produce powerful arguments in favor of evidence-based policies for social change.

A last question for both of you: how can institutions you head, AFD and IRD, coordinate their efforts to mainstream relevant research in the Global South useful to and used for policymaking and implementation?

Rémy Rioux: We believe that the win–win situation is when research is conducted in partnership with policy-makers and civil society. These partnerships need to preserve the independence and rigor of the research and ensure cross-fertilization for the intellectual output to be relevant and contributory to progress in our societies. This involvement should be seen at all stages of the scientific production cycle from the design and framing of the research to its implementation and the production and dissemination of knowledge.

The Intergovernmental Panel on Climate Change is a good example of this momentum. It is essential to combine climate change action with policies that reduce inequalities and strengthen the social link of our societies to ensure that the environmental transition so vital today is also socially sustainable. This much-needed research should be based in the South with the support and assistance of centres of excellence in the North, where appropriate. The South will inspire the North, to quote the president of UNICEF France, Jean-Marie Dru.

In this regard, IRD is an exceptional partner in view of its scientific excellence and close links with academic teams in the countries of the Global South, and because its mandate is given over entirely to the developing countries and all its work is conducted in partnership with and to build research capacities in the South. Our two institutions have been working together on sustainable development research since 2012. Our work together takes a sustainability science approach, promoting interdisciplinary work, and a bridge between scientific knowledge and the knowledge of the other development players.

In the same vein, the International Development Finance Club (IDFC), which I have chaired since 2017, contributes to mainstreaming research in the Global South used for policymaking and implementation. In fact, the Club recently presented a groundbreaking report, produced by independent think tanks CPI and I4CE, which provides a robust framework usable by the 26 national and regional IDFC member development banks—including many based in the Global South—and the financial community at large, to align any financial institution's vision with the Paris Agreement at country, strategic and operational levels. Also,

project evaluation is a subject of discussion within the Club to identify issues raised by the evaluation of climate action by international organizations and donors, including examining methodological challenges raised by the measurement of the impacts of climate change development programmes. In the coming months, the operationalization of IDFC's e-platform will enable the Club to better connect experts with each other and promote the sharing of knowledge and good practices.

Jean-Paul Moatti: The French National Research Institute for Sustainable Development (IRD) has a long history, spanning 75 years in 2019. The institute works in over 50 developing countries. AFD has benefited from the expertise of IRD's researchers for many years. Paradoxically, however, although France remains the only advanced country with an interdisciplinary public research organization like IRD, whose unique mandate is fair scientific cooperation with academics from developing countries, we have worked less frequently and systematically with AFD than other development agencies and banks, such as USAID and the UK's DFID, have with their national researchers. One reason for this is that, until recently, funding research was not part of AFD's mandate, so scientific contributions had to use contractual expertise channels not always well suited to research projects. However, things are changing fast with the SDGs and climate change action becoming a common focus for both AFD and IRD, and scientific diplomacy increasingly recognized as a major contributor to sustainable development (see the IPCC for the climate and IPBES for biodiversity).

First and foremost, AFD now supports programmes to build university and research capacities in developing countries. Examples of this can be seen in Côte d'Ivoire in association with the French debt relief programme and in AFD's work with the World Bank to support African Centres of Excellence in research (ACE), especially in French-speaking Africa. Many of IRD's African ACE project partners now receive this type of support from AFD. Second, the new global agreement between AFD and IRD signed in early 2019 seeks to more closely associate decentralized initiatives between AFD and researchers at country level with major programmes co-developed by the two organizations. These programmes will include scientific evaluation and development projects funded by AFD, which could make for useful learnings and a certain degree of international generalization. The fact of building an experimental or quasi-experimental evaluation protocol ex ante, randomized or otherwise, at the same time as the process of setting up a sustainable development project increases the chances of success for both the scientific assessment and the project itself. So closer collaboration with AFD could be an excellent way to promote the changes in research practices that sustainability science needs, including co-constructing research programmes with the communities and vulnerable population groups directly concerned.

Interview with Gulzar Natarajan

Gulzar Natarajan, during your career as a senior Indian government official, you have accumulated a wide range of experience. You have served in the office of the Prime Minister of India, you have managed the Infrastructure Corporation of the Andhra Pradesh state, you have been District Collector of Hyderabad, you have been Chairman and Managing Director of a power distribution company based at Visakhapatnam, you have been Municipal Commissioner of Vijayawada, and in development field postings across Andhra Pradesh. You have led the design and implementation of large-scale projects in many fields, such as infrastructure, urban, health, education, skills and livelihoods, poverty reduction.

These different positions, combined with your training both in engineering and development studies, give you a sharp look at the question of how Indian bureaucracies work and at development policies, how they can be improved and the type of research methodology that can contribute to this improvement.

What are, according to you, the most important policy issues India is facing today that could be informed by sound research, starting with the macrolevel?

At the outset, let me clarify that what follows are my personal views. There are several issues of critical importance which policy makers in India grapple with on a regular basis. All of them could benefit with insights and evidence from high-quality research in that area. I have myself engaged with several of these issues at some time or other and have been frustrated at the lack of sufficient research which could have shed more light on my working assumptions and on the solutions that were proposed. I have listed twelve areas that I believe are essential: macroeconomy, financial markets, infrastructure, banking sector, industrial policy, public finance, labor maket reforms, informal markets, urbanization, development, foreign and trade policy, and data analytics. I list in Annex A the specific subquestions that I consider essential for each of these questions.

Apart from these macro issues, could you provide some relevant concrete examples focused on more practical issues, real-world Mayors or municipal commissioners have to address in their jurisdictions?

There are three levels of sub-national governments in India—the local government (urban or rural), the district administration, and the state government department. Bureaucrats and political leaders at these levels grapple with multiple challenges, even as they pursue decisions in real-time based on limited available data and insights.

What are the issues, both strategic and operational, that typically agitate the minds of bureaucratic leaders at each of these three levels? What are the technically sound, administratively feasible, and politically acceptable solution choices for each of those issues? How can high-quality academic research inform and

enrich the bureaucrats with their decision-making and implementation? What research techniques are appropriate to support this process?

To give a sense of the challenges, let me outline the super-set of issues that a typical bureaucrat across each of the three government levels in any Indian state is confronted with. While it is not exhaustive, it also tries to cover as many of the broad areas of likely engagement at each level. Given the size of the country, most of these administrative levels cover at least a million people, and state level involves tens of millions. Also in terms of impact, each of these are issues which span across the particular system, and a change from the business as usual can have significant, often transformational, effects. So, these are indeed first-order development challenges with significant impact.

Besides, as a best practice of engaging with a problem, these questions define the common problems and provide a good starting point for research enquiry.

Annex B provides for a listing of the issues that agitate the minds of officials at each of these three levels.

It is evident that exploring answers to many of these questions are not amenable to any one particular research methodology. In fact, the majority of them are not perhaps amenable to neat quantitative approaches, and would need qualitative and ethnographic analysis. A consultant, equipped with rigorous enough toolkits, may be best positioned to explore these. In other cases, a combination of techniques ranging from data analytics to econometric techniques and field experiments may be required.

As can be seen, very few of these questions are amenable to a rigorous RCT. For a start, in most cases, the bureaucrat does not have the flexibility to create treatment and controls. Second, isolating a problem and its potential solutions neatly, so as to explore attribution, is most often impossible. Third, being embedded in large and complex systems, where the influencing factors are not easily identified, contributions rather than attributions are easier to explore. Four, the bureaucrats take decisions in real time, and therefore do not have the luxury of long-drawn experiments. Five, most often the immediate results of these interventions merely reflect partial equilibriums, and steady-state results take a long time to become evident. So headline evaluations rarely serve the purpose. Finally, related to the previous point, there is no one good solution to a problem that can be picked up and implemented. Instead, problems (including internal sabotage) start to surface when the solution hits the road, necessitating iterative adaptation, especially in the initial period, before it stabilizes.

In terms of field experiments, more than the long-drawn RCTs, a more relevant technique may be quick and dirty A/B testing which can give insights about competing choices evident at decision-forks during implementation. The idea is to check whether certain proximate indicators of likely success (identified from an examination of the theory of change for the intervention) are being adhered to. In the case of such complex issues and challenging environments, and where outputs

and outcomes take time to surface, verifying compliance with proximate process indicators may be more relevant and practical than headline evaluations for outputs or outcomes.

For the kind of questions that RCT proponents raise, do they bring the right answers and how far their answers can be useful?

I will respond with three specific examples. Consider the case of the fight against driving drunken driving, which has been studied by A. Banerjee and colleagues (Banerjee et al. 2012) and summarized in a press article (Banerjee et al. 2017c). They argue in favor of the increased use of breathalyzers (tools) and introduction of higher fines (laws) in the fight against drunken driving. This, they also conclude, has to be complemented with a strategy of vehicle checks at randomly decided locations using dedicated teams of police drawn from the reserves.

There is nothing new about this "strategy." It is commonly used by police superintendents and commissioners for short durations when something (usually a high-profile accident or a court directive) triggers greater vigilance on drunken driving. The problem is that these things cannot go beyond short periods.

The challenge, as Esther Duflo has pointed out in other contexts, is with the plumbing. Take the issue of random locations. While, from the outside, allocating random locations may appear an algorithmic exercise, it can be challenging to operationalize at scale. A combination of closing ranks by powerful and entrenched interests and systems with very weak institutional capacity, especially at the police station level, mean that such strategies are easily compromised or diluted, unless there are deeply committed leaders micro-managing the process. As an example, random third-party quality checks of engineering works under construction, now commonplace, is frequently compromised by collusion. As to reserves, their deputation beyond a few days is impossible. In a heavily under-staffed police force overburdened with crowd and VIP events management responsibilities (bandobasth), reserves are that only in name. The competing demands on reserves are too many.

In fact, the evidentiary standard for a legal offence of drunken driving may limit the role for reserves. For example, in order to limit discretionary excesses, the law (regulations issued on the central Act in different states) mandate that breathalyzer tests have to be carried out in the presence of a police personnel above a certain rank. And over-burdened policy administrations have too few personnel of such rank to spare for any traffic responsibilities, much less for night-time drunken driving patrols. But delegating this responsibility raises legal and practical problems.

The above findings came from an RCT conducted in 162 police stations in Rajasthan covering five management interventions: limitations of arbitrary transfers, rotation of duty assignments and days off, increased community involvement, on-duty training, and "decoy" visits by field officers posing as citizens. It found that

the first three "which would have reduced middle managers' autonomy, were poorly implemented and ineffective," while the last two had "robust impacts." Based on these findings, the researchers found "very large outcomes" from an intervention that linked "good performance" on sobriety tests *without relying on middle managers* to the "promise of a transfer from the reserve barracks to a desirable police station posting."

The paper claims, "The experimental results in this paper show that it is possible to affect the behavior of the police in a relatively short period of time, using a simple and affordable set of interventions." This claim is misleading.

Let us examine the "strategies" that the researchers have explored. We have already examined the challenge with random inspections and drawing from reserves. "On-duty trainings" are a staple of any administrative system. The problem is just that the trainings do not get translated into meaningful enough learning or internalization. This, by the way, is a common failing of in-service trainings for teacher, doctors, inspectors, and so on.

"Decoy" visits by field officers is hardly a new idea. It is a staple of human intelligence gathering (humint). Again there are practical difficulties in identifying and managing their activities. Further, such decoys also have the potential to create systemic distortions that do more harm than good. Therefore, rather than focusing on piece-meal fancy measures like decoys, police managers should expend efforts in improving their intelligence wings and special branches and using multiple channels like decoys, third party agencies, and soliciting telephone feedback from complainants. In fact, numerous police leaders across India routinely variations of such approaches to obtain feedback.

Take the case of the last intervention, linking transfers of reservists to good performance and reducing the autonomy of middle-managers. Again, plumbing is the challenge. For a start, drunken driving enforcement often does not figure among the top priorities of the mainstream police force, leave aside the reserve force. And their (reservists') main bandobasth activities are not amenable to quantitative assessments of individual policemen. Second, how sustainable is a policy that explicitly seek to reward some policemen with transfers to "desirable" places (for whatever "good performance") and penalize some others (the natural corollary) by drafting them to reserves?

Third, the moment we start linking incentives to quantitative performance indicators in detecting drunken driving, it is only time before we get into targets and slip down an undesirable slope. This is a common feature of public systems which seek performance measures linked to high-stakes personnel management decisions. Fourth, what is the sustainability of an administrative process where there is no involvement of middle-managers? Ultimately managers at some level have to be managing this institutionally. Even assuming that level exists outside the "middle-managers," are we any less likely to have concerns with them? And is it even practical to think about such administrative activities without the

involvement of middle managers? Finally, it is far from true that "evaluation generated evidence and information is not typically available to the police leadership." "Evidence," of a far more sweeping breadth and with more than the requisite credibility and rigor, in large measure, is available, to police leaders who keep their eyes and ears open. No amount of careful quantitative evidence generation can get you beyond a few baby steps in the endeavor to effectively manage large systems.

The researchers point out that successive Police Reform Commissions have not only not advocated the three "successful" interventions but also recommended the "ineffective" interventions. For a start, Reform Commissions recommend on institutional reforms like limits on transfers, community involvement etc., and avoid routine and commonplace operational measures like use of decoys or on-duty trainings, much less impractical ones like performance management of reservists. Further, these recommendations are essential plumbing necessities of any well-governed administrative system. In contrast, the researchers' solutions, as discussed above, suffer from serious practical deficiencies. The authors of the Reform Commissions refrained from such band-aid recommendations because being life-long plumbers, who, with varying degrees of success (or failures), had grappled with the plumbing challenges of policing in the real world, they were responsible and honest to not do so.

The point I am making is that these ideas—random inspections for drunken driving, limits on arbitrary transfers, rotation of duty assignments and days off for police personnel, increased community involvement, on-duty training, and "decoy" visits by field officers posing as citizens—are all good, and don't need any evidence of proof. It is not lack of evidence that is in the way of their adoption. But implementing them at scale is hard, and depends on the interest and commitment of the police leader concerned, and some of them require the sort of resources/capacity which the system currently does not possess. In weak state capacity systems, such interventions run on individual (the leaders) and not on institutions.

Let me take another example: an RCT conducted on third party audits of polluting industrial units in Gujarat, published as a paper (Duflo et al. 2013) and as a policy brief (J-PAL 2013). The authors claim to provide evidence that independent third-party audits are effective at reducing environmental pollution.

In brief, in response to a High Court directive, certain types of highly polluting industrial units in Gujarat had commissioned and had been filing thrice-a-year third party audit reports since 1996. But its performance was less than satisfactory. The researchers found that instead of being paid by the firms themselves, once the auditor payments were made from a central pool, audits were conducted randomly, and auditors were incentivized with a bonus for accuracy, there was a significant increase in reporting of pollution readings and reduction in actual pollution itself. In order to strengthen their theory of change, the researchers also

conducted explicitly announced random sample back-check super audits of each auditor and their payments were made contingent on the accuracy of the original audits (compared to the back-check super audits).

Now, no one disputes the value of third-party audits done through independent agencies, paid from a central pool, and reinforced with back-check super-audits. Random sample (and unannounced) third-party audits or certifications are today commonplace in monitoring everything from engineering works to the quality of goods procured and services delivered. In India, over the last couple of decades, third party quality audits, ostensibly of random samples and carried out unannounced, have come to be embraced for engineering works executed by all departments, big and small, urban and rural. It has undoubtedly contributed to improving the quality of these works, and where done well the benefits are very significant. And perceptive environmental protection officials across states are aware of its utility.

So, did the research provide anything that was valuable for Pollution Control Boards (PCBs) across India? In the real-world of weak state capacity, effective management of third-party audits itself is a massive task. The back-check super audits make the task even more onerous. Since the back-checks were conducted under the supervision of enthusiastic and committed research associates, were known to the industries, and auditor payments were made contingent on original audits tallying with the super audits, the original audits could not but not have become high stakes and thereby also of high quality (Hawthorne effect). The RCT established the efficacy of this particular double-audit design.

Unfortunately, such a two-level audit, which achieves both high stakes and high quality in this tight manner, while desirable, is too high-stakes and engagement intensive to stand a chance of effective scale-up through weak public systems and where pollution is the norm than the exception.

The researchers found at least two channels of incentive distortion with the earlier approach. One, an agency problem arising from the auditors being paid by the audited. Two, the auditors being paid significantly less than would have been required to conducted good audits.

On the agency problem with how the auditors are paid, the researchers need not have taken the trouble of an RCT since there is a very rich body of literature on the problems with credit ratings shopping by financial institutions. The idea of not having auditor/rater payment being done by the audited/rated is widely accepted. Second, the auditors are being paid less because these audits had become largely a proforma exercise, and all sides know it.

It is difficult to believe that the Gujarat Pollution Control Board (PCB) did not know what was going on. I can think of at least five first-order plumbing reasons why the PCB preferred to go along with status quo than adopt what were obvious (if at least because they were, as mentioned earlier, already being done in other

sectors) reforms. One, these reports were being generated at the instance of the High Court and being reported to them. As long as the Court was happy, there was no reason nor inherent motivation for the PCB to change the system. Such proforma compliance with regulatory requirements is not uncommon. Two, if the government decided to do the audits, there was the question of who would pay for it or how the amounts would be collected, not to mention the "headache" of managing this additional administrative responsibility. Three, there are strong vested interests among the polluting industries that prefer the status quo. And regulatory capture and administrative tolerance, especially in such, is always imminent. The difference between business as usual and pollution compliance for these firms is in many cases a matter of survival itself, and with large employment implications. Four, a step change with effluent emissions by tightening standards abruptly and rigorously would only force the close-down of several units. And, as seen from numerous precedents, this goes against a very sensitive political economy. Finally, whether we like it or not, pollution control has been, for long, a marginal concern for most state governments as they chase economic growth and job creation. Accordingly, the resolve to undertake pain-staking reforms has been limited.

Let me take a final example. Karthik Muralidharan and his colleagues conducted an RCT on the use of mobile phones to improve governance (Muralidharan et al. 2018a, 2018b). In brief, telephone calls based feedback was elicited on the quality of implementation of the Telangana government's Rythu Bandu Scheme where direct cash transfers through checks were made to eligible farmers—did farmers get the check, did they get it in time, did they encash etc. An RCT evaluation of the telephone calls revealed that 83 percent of farmers received and encashed their checks, farmers in areas with such monitoring were 1.5 percentage points more likely to receive and enchase their checks, and among the bottom quartile land holding farmers percentage was 3.3 percentage points higher. The call centers, the authors claimed, delivered an additional Rs 7 Crores (Rs 70 million, ie US$1 million) to farmers at a cost of Rs 25 lakhs (Rs 2.5 million, i.e. around $US 35,000).

Again the point is: Who disputes the efficacy of this idea? Did this require an RCT? The practice of soliciting feedback from citizens and customers through telephone calls has been in vogue for several years. Several public agencies across states have had citizen feedback electing mechanisms in place for years. Neighboring state of Andhra Pradesh has even made this central to performance assessments by evaluating most government activities by telephone feedback through its elaborate Real Time Governance System (RTGS). Since a long time, several power distribution companies and municipal corporations have had such telephone call centers to elicit feedback.

No bureaucrat would dispute the underlying idea—citizen feedback using random sample telephone calls is a useful way to assess implementation quality. This raises two issues. One, is this the most sustainable and cost-effective

approach to enhancing implementation quality? For instance, as we shall discuss latter, improving the existing monitoring system would have come without any additional cost and with other benefits.

Two, even if we pursue telephone feedback, what about the problems that follow? For a start, with a telephone feedback system, the real challenge is not about getting feedback, even granular and actionable feedback, but the ability of the system to act on any feedback in a meaningful enough manner. That is the real binding constraint and that is critically dependent on the state's capacity to engage actively on a basic governance issue—monitor and act effectively on information. Furthermore, we should not discount the scale scenario where such monitoring, without the follow-up requirements, is most likely to become one more addition to the monitoring paraphernalia without any incremental benefit, but at a significant additional cost. Besides, managing the telephone call centers and the feedback management system itself takes up considerable scarce administrative bandwidth. Finally, this can distort the incentives within the bureaucracy and weaken the existing monitoring mechanisms.

In fact, one could very easily imagine a scenario with such ideas. Telephone feedback systems can improve implementation efficiency, and they look different and appealing. So, let's establish telephone call centres in each district/state. And in five years, we could have another addition to the graveyard of development innovations, dysfunctional call centres and vast amounts of money down the drain.

Incidentally, I have personally been engaged with the implementation of all the three interventions discussed in the papers above. I have grappled with all the messiness and practical challenges associated with implementation. Third party audits and telephone feedback have been a constant across multiple postings since at least 2005.

Keeping with RCTs focus, what would be your own answers and how they compare to their suggestions (for instance in the field of governance, to improve the Indian administration performance)?

Let's start with the policing case. Instead of suggesting solutions that help improve institutions and systems in which these policing activities are embedded, the researchers end up recommending piece-meal and unsustainable solutions which do not sufficiently account for contextual factors and systemic considerations.

Regarding the prevention of drunken driving, the objective of research, for example, should have been to improve policing outcomes by *enhancing accountability of middle managers* and *figuring out sustainable institutional solutions*, instead of *dispensing with middle managers and ad-hoc hiring of decoys*.

Performance incentives that involve large enough financial rewards or transfers are unlikely to work at scale in complex public systems. After all, where does it work in even developed countries? For a start, quantifying outcomes in a

credible manner is deeply problematic, and collecting and managing it even more so, in most such contexts. Some limited evidence of success with monitorable logistical activities like tax collections does not mean the same can be applied with similar expectations for teachers or police. Second, over time, financial incentives most likely end up becoming entitlements, thereby worsening the problem of already high lower level government salaries. Finally, postings and transfers are among the highest stake administrative actions, and when done in environments where rank ordered "good performance" cannot be credibly and indisputably established, it can be a recipe for controversies and discontent.

There are very few innovations, either great ideas or process re-engineering or even management theories, that can sustainably and meaningfully "affect the behavior of the police in a relatively short period of time" in conditions of acute systemic and leadership weaknesses. All things being equal, wherever police systems work well, it is more likely a combination of functional administrative capacity and good leadership. The intensity of the latter can even temporarily mask deficiencies in the former. It is for this reason that we keep hearing stories of poorly run administrations suddenly becoming efficient with the arrival of a good Police Commissioner, and returning to status-quo-ante when they depart.

In economics-speak, the production function for good policing outcomes is largely these two. Innovations most often can work at the margins to improve administrative capacity and free up leadership energies for productive use elsewhere. But in really weak systems, as we have here, leadership is necessary to both generate short-term good outcomes and build long-term institutional capacity.

Addressing this state capacity enhancement challenge was an opportunity that the researchers engaged with the Rythu Bandhu scheme in Telangana had. The underlying problem for the researchers was to enhance the effectiveness of implementation of the scheme. The telephone-feedback system is just one approach, one with several potential dangers. Instead, the researchers could have used the opportunity to figure out how the monitoring of implementation could be improved. This would have been useful to unpack state capacity issues, especially the critical issue of more effective use of supervisory systems to monitor the quality of program implementation. Unfortunately, these are not amenable to RCTs.

What is a more sustainable and effective monitoring approach to review development programs? Consider some (there could be more) of the variables associated with such reviews—who is reviewing and who is being reviewed, frequency of review, specific parameters being reviewed. So, a District Collector (or a Block Development Officer, BDO) could once a week (or once a fortnight) review BDO (or Village Revenue Officer, VRO) on some process (approvals processing) or output (receipt or encashment) indicators.

How about optimizing this? Let's say all blocks divided into two treatment arms and two different review methods of/by BDOs or VROs (with a none-too-onerous deep-dive you can figure out these options), against the business as usual monitoring control group.

This would immediately spotlight attention on the importance of the quality of monitoring and the role of state capacity improvements (in a more objective manner the likes of which has never been done before by researchers) in effective implementation. For example, if it shows that BDOs who review VROs once a week as against those who review once a month are associated with x% higher cash transfer receipts for the poorest quartile, then that creates a trigger with a meaningful and actionable insight for the Secretary to Government of Telangana state who is implementing the Rythu Bandhu scheme. In fact, it would have actually delivered more returns at virtually no cost.

Further, the gains go beyond just improving the efficiency of the specific intervention. It would likely apply to most interventions implemented in that jurisdiction. It would have genuinely spotlighted attention on state capacity, specifically how better monitoring would have increased the efficiency of public service delivery. This may appear self-evident, but in a world where everyone is searching for innovations and different ways of doing things, what should be so obvious often actually ends up being marginalized!

In the same vein, the issues of relevance for practitioners for third party audits are more in the plumbing. What would have been of great value for interested PCB officials is the design of a robust independent third-party pollution audit system. There at least a few that come to mind immediately. What should be the most cost-effective design of third-party audits? What number, frequency, and scope of inspections would be most cost-effective? What can be done to mitigate the risk of capture of such audits? How can the audits respond to dynamic expectations? How would the inspection patterns have to change in response to the strong likelihood of adaptation and gaming of the audits by the industrial units? And, how can the audits be sustainably financed?

In case of an engineering work, the most cost-effective design would focus on the least number of samples with the longest periodicity that would not compromise on deterrence. As regards addressing capture, inspections may not only have to be random but also done by personnel on rotation, and the agencies themselves may have to be shuffled periodically or multiple agencies employed. As to dynamic expectations, it may be necessary to periodically revisit the audit criteria and calibrate for the adaptations. On financing, it may be required for the PCBs to commission and finance the audits and, maybe, recover a part of the cost as a user-fee (or from the fines collected, though it could have perverse incentives) paid into a common pool. These are the messy details of implementation that concern the bureaucrat.

Ultimately what practitioners need is an administratively simple and workable third-party audit design. Or more specifically, they would need a tender document

that captures these design specifications. The aforesaid research does not provide anything of relevance on this.

Another recommendation, in relation to your second question, is to considerably broaden the scope of the research questions to cover the list I proposed to you earlier.

What would be your takeaways for researchers from these examples?

At the outset, let me be clear that my purpose is not to downplay the importance of RCTs or field experiments in such areas. They have undoubted value, but have to be seen in the true perspective. There will always be instances where governments will be faced with taking a decision on competing choices. They help fortify the case for compelling ideas, contributing to building up the momentum for their scale-adoption. Equally important, they help with generating the evidence that can help the push back against bad ideas. At the least, it provides some basis for a government official sitting on the fence to bite the bullet with independent and/or random sampled audits. Fundamentally, we need the full toolbox of qualitative and quantitative assessments to help generate insights to design and improve implementation of development interventions.

All the three ideas share an important feature, and this is characteristic of many RCT studies. These are all stand-alone technical fixes with a logically neat appeal when seen in isolation. The decoys are simple and nifty; independent third-party audits paid from a central pool and validated with back-checks are logically water-tight; and call-centres and telephone feedback convey the form of independence and simplicity. All three appear new or innovative, in that they are not the norm. They have an irresistible appeal when seen against the problem in isolation and the dismal landscape of failures of the regular administrative responses. And it is possible to run short-duration pilots in all these supervised by high quality research assistants and generate evidence of efficacy.

Unfortunately, when the rubber hits the road in scale implementation, all these appealing features ironically end up being among its failings. Several factors, which were overlooked, start to bind. The state's capacity to administer and monitor, masked by the small size and presence of energetic research assistants in the field experiment, gets exposed. Logic gets torn apart when faced with practical challenges. The system takes over.

Most bureaucrats know about these and the perceptive ones refrain from such piece-meal and band-aid type solutions, and adopt them as part of systemic reform efforts. Officials seeking popularity or quick-fixes have a preference for such piece-meal solutions, and they invariably wind-up after their transfer.

In the three cases, evidence generation from research has an important role to play in improving policing, designing a robust third-party audit system, and rigorous monitoring of scheme implementation respectively. The answer to the bureaucrat's problem statement was not a headline RCT evaluation, involving

difficult to implement innovations. Instead, it required more sustainable institutional solutions.

After a comprehensive problem-solving, the specific enquiry should have been something like this—conditional on the efficacy of random surprise inspections for drunken driving, third party audits to verify effluent emissions, and monitoring systems to enhance effectiveness of Rythu Bandhu scheme implementation, what should be the most sustainable, practical, and cost-effective approach?

These, as mentioned earlier, would have required a more heterodox set of toolkits or a combination of quantitative and qualitative methods. It would have involved short-duration A/B testing to figure out uncertain elements, ethnographies/qualitative studies to identify critical processes etc. In the course of implementation, there may also arise the need to do a full RCT evaluation between competing program design alternatives.

Further, we should note that there is nothing Gujarat or even India-specific to many of these plumbing issues. The broad contexts are the same—weak state capacity, centralized bureaucracies marked by low trust, scarce resources, overburdened bureaucrats, and challenging work environments. Political economy factors are another layer of complication. They are universal to most developing counties. So many of the plumbing features that could be tested have generalizable features. These research endeavors therefore represents perhaps missed opportunities, and more worryingly, many researchers may not even be aware that those outlined earlier are the real challenges.

These arguments are motivated by my strong belief, from observing the origins of numerous such studies and even heading government agencies in places where several such experiments (not these three) were being undertaken, that field research rarely ever start with a felt-need or felt-problem of the government officials. Most often the Principal Investigators come across an idea, mostly in the form of a stand-alone intervention, with some convenient rationale for engagement, which they work backwards to develop into a solution hypothesis, secures funding from a donor, and then approaches the government interlocutor with an evaluation proposal. It is unlikely that the government official will have a problem. But neither does he or she have much of a stake in the result. As a counterpoint, governments all along hire consultants to solve specific problems. There is a skin-in-the-game associated with these engagements. So the engagement starts with a deep-dive of the problem context and solution alternatives emerge (problem-solving is more comprehensive), even if the solution analysis is less than rigorous, and the results get taken seriously, whether implemented or not.

While there are also practical difficulties, inadequate comprehension of the real plumbing challenges among researchers is the bigger obstacle to engaging directly with the problem. It does appear that the best plumbers, outside of actual

plumbers, are the practitioners themselves. Plumbing insights is more a lived experience than a learnt theoretical knowledge.

Annex A. A List of Research Agenda on Indian Economy

It is a matter of concern that important issues related to India attract limited research interest. Instead, the only area of interest concerning India among top researchers revolves around poverty studies of the kind involving randomized control trial (RCTs) and romanticized visions of entrepreneurship and social enterprises that serve the Bottom of Pyramid.

Research can range the spectrum from econometric analysis to ethnographies and event/case studies. The objective should be to promote the highest quality of research which can reliably inform the debates on those issues in India and thereby influence policy-making in those areas.

An illustrative list of topics on which high-quality research can contribute significantly to improving public policy in those areas is as follows.

1. Financial markets
   a. Monetary policy—How does India's version of heterodox monetary policy leading upto and during the Global Financial Crisis compare to the orthodoxy? Assessment of monetary policy transmission, its quantification, likely constraints, and comparison with other countries? Assessment of the recent drivers of inflation in India? Is inflation targeting an appropriate monetary policy framework for India, or should India embrace a more heterodox set of tools, and if so what should be its components?
   b. What has been India's experience with its capital flows management measures, especially compared to its peers? How has India managed to mitigate the effects of spillovers from sudden stops and capital flights? What are the lessons from exchange rate management policies?
   c. Capital markets—How have the different parts of India's capital markets evolved over time in comparison with peers? What is the level of global integration of different segments of the domestic financial markets? How does India's capital market regulation compare with those its peers?
2. Infrastructure
   d. What has been India's experience with private participation in infrastructure compared to those of Latin American and European countries—evolution of PPPs, trends on cost and time over-runs, Value for Money (VfM) assessment and Public Sector Comparators (PSC) with respect to

PPPs, problems of aggressive bids and reckless financing, renegotiations and its frameworks, contract/concession variants, financing sources, asset monetization etc?

e. How do India's infrastructure companies, their business practices, and their financing strategies compare with those of other major economies (especially European)?

f. How do the costing schedules and contracting practices/approaches of public procurements of infrastructure in India, China, and elsewhere compare?

3. Banking sector

g. How does India's experience with periodic banking sector crises and its resolution compare with that of Sweden, US, Spain, Ireland, Iceland, Italy, etc?

h. How do the management practices, internal controls, and credit appraisal processes of public and private banks compare? How does government's micro-management distort incentives within public sector banks?

4. Industrial policy

i. What is the assessment/quantification of India's SEZs in job creation, output increase, and the broad area of externalities including technology spill-overs and displacement effects? How does it compare with China's SEZ policy? How could a Charter Cities based (or rules-based) approach be differential, and what are its likely quantified gains?

j. What is the assessment/quantification of India's tax and inputs concessions based industrial policy, both in terms of direct benefits and externalities? What has been its impact on the growth of SMEs compared to the larger enterprises? What are the alternative industrial policy levers that can be used, from experience of other countries, and its likely impact on tax revenues and economic growth?

k. What has been the relative impact of the primary industrial policy levers of fiscal concessions and input subsidies on SMEs and large firms?

l. An assessment of the EoDB reforms and comparative studies—what has worked and what has not?

m. What has been the relative net job creation by large enterprises and SMEs, impact of MNCs on domestic enterprises and their productivity, technology transfer gains from MNCs? What are the relative impacts on productivity, output, and job creation from FDI and domestic investments?

n. Why does the Global Value Chain elude India? What can be done to connect India into the GVC? How have countries established and tightened connections with the GVC and learnings for India?

5. Public finance
   o. Assessment of India's fiscal federalism compared to that of other major democracies?
   p. Assessment of India's direct and indirect taxation system on various dimensions—descriptives and comparisons, response/elasticity to various fiscal policy instruments and growth indicators, impact of informality on tax revenues, incentive distortions arising from aggressive enforcement, etc?
   q. Evolution of India's subsidy regime—its relative efficiency and effectiveness over the years and compared to others?
   r. How do the different GST regime and rate alternatives compare with each other in terms of revenues mobilization, business profits, and economic growth?
6. Factor market reforms
   s. What has been the impact of the critical provisions of the Industrial Disputes Act, especially those related to exit, and the relative performance of UP and other states which have had higher thresholds?
   t. What have been the costs associated with the multiplicity of labor regulations and attendant compliance/filing/reporting requirements? What has been the primary deterrent to bringing together thousands of employees under a single umbrella, especially in sectors like textiles?
   u. What are the biggest sources of land market distortions, what have been their respective effects, and what policy options are available to remove the distortions and how do they compare?
7. Informal markets
   v. A descriptive study of India's informal economy and its positive and negative externalities. Its evolution since the early nineties liberalization and comparison with that of peers? Insights about how to engage with the informal economy and reduce its role—should it be more by forcing firms go formal or encouraging new firms to start formal?
   w. How do formal and informal markets interact with each other? What is the impact of large firms on the productivity and growth of firms in the informal markets?
8. Urbanization
   x. What are the costs associated with India's low FAR and consequent urban sprawls, in terms of housing affordability, urban commutes, and environmental pollution? What is the cost on urban productivity and growth due to low FAR?
   y. An assessment of property tax revenue collections compared to peers? What are the possible sources of local government revenues in India and their respective potential? What is the potential for value capture finance and FAR purchases?

z. The magnitude and scale of India's affordable housing problem and its economic consequences. What are the policy responses from the experience of other countries? Has urban renewal been accompanied by gentrification and its impact on inclusiveness? What are the important drivers of urban property prices? How has India's urban public housing program worked—its impact and comparison with other countries? The relative impacts of the major instruments—mandates, higher FSI, public housing, release of public lands—on housing prices and affordable housing stock.
   aa. Assessment of India's urban utilities provisioning compared to peers? What has been the cost of externalities due to traffic congestion, air pollution, interrupted water supply, deficient sewerage facilities and discharge into rivers, open solid waste dumping etc.?
   bb. How does India's property tax system compare with those of its peers? How has it responded to various policy instruments and infrastructure augmentation? How do tax rates vary across Indian cities, and what learnings about good practices and positive trends from them?

Annex B. Issues of concern to policymakers

How many of these questions are amenable to being answered by RCTs? Instead, how many of these could have been answered by rigorous ethnographies, comparative studies, and other forms of research?

1. Issues of concern—Municipal Commissioner/Mayor—all cases, cost-effective, politically and administratively possible
   a. What would be the most efficient, least distortionary, and simplest tax slabs for property taxes?
   b. What innovations can reduce underassessment by tax inspectors—self-certification, internal random inspections, outsourced random inspections, periodic re-surveys, technology? What would be the most efficient delegation of powers for tax assessments?
   c. What is the most appropriate incentive mechanism for tax inspectors to maximize bill collections?
   d. What approaches are possible to improve tax collection efficiency—rewards, punishments, nudges, bundling, community mobilization, shaming etc? Which are the most cost-effective ones?
   e. How to ease political opposition to imposing new categories of taxes—door-to-door garbage collection, road cess on fuels, congestion charges etc—bundling, opt-out etc?

f. How to streamline and simplify buildings approval process—where self-certification can be made sufficient and where not? What process innovation can control building violations? What would be the most effective and least distortionary delegation of powers for buildings approvals?
g. How can pilfered water, sewerage, and electricity connections and unassessed properties be deterred? What are the most effective ways to deter such pilferage?
h. Massive amounts have been invested in water and sewer networks, but only a small proportion of people end up taking connections due to high connection costs and other barriers. How do we ensure that households take water and sewer connections once the network is put in place, without compromising on the revenue recovery? How can the access barriers be mitigated?
i. What should be the most practical (not necessarily technical optimum) preventive maintenance schedules for drains, sewer networks, motors, roads, street lights, etc?
j. How to reduce traffic congestion, especially at specific road stretches and specific times during the day? Or how to reduce congestion during school opening and closing times?
k. What is the best possible strategy for the adoption of a phased transit oriented planning—higher FAR around which transit stations or certain specific locations?
l. How do I increase the use of public transport? How do I discourage the use of private vehicles? How do I encourage a greater number of people to use public transport?
m. What should be the most effective strategy for allotting the small number of public housing constructed each year among several-fold competing applications—lottery, criterion, etc? How to ensure that they do not sell it off and squat again in slums?
n. What would be the most effective strategy for deployment of sanitation workers—street/drain lengths, garbage pick-up locations, etc?
o. What should be the most effective inventory management schedule for the municipal stores—sanitation supplies, street light consumables, utility network spares, etc?
p. Most effective (and least distortionary) inspection schedule design for municipal bill collectors and sanitation workers?
q. What is the most appropriate strategy to reward bill collectors, sanitation workers, building inspectors, etc?
r. What are the most effective strategies to minimize littering? We have placed several litter bins in public places, still people not using them. So, what should be done?

s. How to prevent open defecation? How to prevent street urination? How to ensure that the constructed public toilet is used?
t. What is the best possible approach to manage public toilets? What strategy maximizes the uptake of community toilets? How can community mobilization be used to enhance uptake?
u. What mechanism to streamline street vendors—regulations, incentives, nudges?
v. What is the most cost-effective design of third party audits of engineering works, teacher/officials attendance, performance of specific activities etc? What is the least size of sample, frequency of visits, and breadth of examination/inspection that is deterrent enough?
w. How to improve highway safety—accident locations, times and causes are highly localized and specific—use nudges, target patrolling, etc?

2. Issues of concern for District Collector/head of county government?
x. What should be my three top priorities in health, education, agriculture, livelihoods, and social protection? What handful of outcomes should I monitor on each, and how to do so?
y. What should be my rural development priorities? What outcomes to monitor, what and how (on each) to do so?
z. What should be my priority in terms of Urban Development and Industrial Development? What handful of outcomes should I monitor on each, and what and how (on each) to do so?
aa. Which are the 5–10 infrastructure projects that I should monitor periodically? What periodicity?
bb. How do I ensure that leakages in the various transfers are minimized? Is there a technology solution? Is there a process re-engineering option? Is there a nudge possible? Are the leakages concentrated in some set of transactions in the long transfers chain?
cc. What are the three best uses for the Rs 100 million annual innovation budget available?
dd. What are the problems with the procurement systems across Departments? How can procurements be made more transparent and cost-effective?
ee. What is the most effective way to review my Departments (and flagship programs)—what should be the periodicity of review of officials at different levels, what should be the review agenda for each level, and how to follow-up effectively?
ff. How can I make the best use of data for monitoring field activities and progress of program implementation?
gg. How do I optimize the effectiveness of my field inspections?
hh. How to minimize unauthorized absence of officials?

ii. The District Collector chairs numerous committees. How do I prioritize work among the different committees?

jj. How do I make my Collectorate grievance redressal system more effective—striking the balance in ensuring that it is truly a last-resort window for citizens and not a first-resort window?

kk. How to most effectively manage third party audits of engineering works, teacher/officials attendance, performance of activities etc? What is the least size of sample, frequency of visits, and breadth of examination/inspection that is deterrent enough?

ll. How do I motivate my employees (change the account/narrative)? How do I extract work from them? How do I empower them to make them own up their work?

mm. What is the right level of delegation of responsibilities to the various departmental heads, one that promotes efficiency and ownership without engendering too many problems?

nn. How to improve learning outcomes? Towards achieving this objective, what are good uses of the Rs XXX available each year for discretionary spending? In general, what are good uses of discretionary spending budgets of different departments?

oo. What should be my signature initiatives, and what should be the basis for identifying them?

pp. What are the beest inspection strategies for health and education officials at different levels? Will a checklist work? What is the most credible and comprehensive enough second-best checklist?

qq. How do we make the most effective use of the field visits of extension officers of agriculture and other services (which offer extension services)?

rr. How should I rationalize my staff deployment—right numbers of people looking after the priorities in that sector?

ss. What is the most effective system for inter-departmental coordination?

tt. How do I manage to mobilize resources so that my resource-constrained field offices have enough money to meet their regular non-salary expenditures?

3. Issues of concern for policy maker—Health department

uu. What is the best use of my $XXX of untied part of health care budget? How should I allocate this budget between primary, secondary, tertiary, medical education, and other activities?

vv. Are my teacher/doctor/personnel resources optimally deployed? What should be the best personnel deployment policy—criterion for transfers—points system? What is the best deployment strategy based

on functional requirements–specialists and basic doctors? How do I prevent ad-hoc transfers/deployments?

ww. What should be the most effective matching and counselling system for medical college and para-medical seats?

xx. What medical regulations need to be reformed and how in the first phase, the ones likely to generate the most obvious and greatest benefits and those which are feasible?

yy. How do we most effectively regulate private medical education institutions and private hospitals? How do we regulate medical practice—what should professional associations do and what governments? How should the powers of regulation be delegated?

zz. How do I integrate private providers into the referral system? Should there be regulation—rate contracting at some regional level with some periodicity?

aaa. How to improve the quality of treatment advise in PHCs? Should treatment protocols be mandated? If so, how to enforce them?

bbb. What is the most efficient data-capture strategy for electronic medical records? Which are the low-hanging fruits for digitization and work-flow automation? What should be the work-flow automation for each initiative?

ccc. How do I improve my epidemic response protocols—what process re-engineering, what degree of automation, what level of delegation?

ddd. Given my resources, what should be my non-commuicable diseases strategy? How much beyond screening should government services engage?

eee. How do I improve the effectiveness of trainings? What should be the periodicity of trainings? What should be the training content? At what geographical jurisdictions would trainings be most effective? Should trainings be supplemented by coaching? Who should be the coaches?

fff. How do I incentivize the nurses/ASHAs for institutional deliveries and immunization—in-kind, points, or financial incentives? If financial, what is the most likely cost-effective amount? Or should I leave the details to states (or districts) and hold them accountable for outcomes linked to an aggregate incentive?

ggg. How do I design the least distortionary and most practical transition from line-item to performance-based budget? What is the best possible strategy to initiate a move towards outcomes-based budgeting from the current line-item based budgeting? Which areas or parts of the budget are most amenable to initiate the process?

hhh. What areas should I promote partnerships with the private sector, and what should be the strategy in each?

iii. How do integrate the less than fully qualified unregistered medical practitioners (or quacks), who form the vast majority of point of first contact, into the system? What should be the design of their training module?

jjj. How do I rationalize the disciplinary proceedings and streamline the court cases related to health department personnel?

kkk. What changes can be made to the recruitment policy so as to make it more effective, transparent, and credible?

lll. What changes should be made to the procurement policy? How do I make it more transparent? What should be the most efficient delegation of procurement powers? How do I more efficiently manage the drugs and consumables supply chain management? What purchases should be delegated to local level?

4. Politician's problems—you get miserable once you start even thinking about these…

mmm. How do we manage a schooling system where teachers are used only for teaching while also running a democracy with census, surveys, elections, recurrent disasters etc which demand their participation?

nnn. How do we run a democratic jurisdictional unit where nobody wants to pay taxes but everyone wants world-class public services?

ooo. How do we regulate contracts and markets with a chronically understaffed and ill-equipped regulatory system?

ppp. How do we manage to do big bang reforms with a state which is so weak that it cannot run mid-day meal kitchens?

qqq. How do we decentralize and empower officials in a system where corruption and perverse incentives are pervasive?

rrr. How do we ease out inefficient and distortionary subsidies and usher in a more efficient and effective program without generating widespread political backlash?

sss. How do we know that cash transfers are a more effective (in terms of achieving its objective) substitute for in-kind transfers?

ttt. How do we do what is best for the country and still manage to win elections?

Interview with Ila Patnaik

Ila Patnaik, you are presently professor at the National Institute of Public Finance and Policy, New Delhi. Prior to this, you were the Principal Economic Advisor to the Government of India. This gives you a dual competence to assess the role of RCTs in terms of policy making in India.

Why has India become such a popular destination of RCTs?

India has practically become a hub for RCT studies among developing countries. The American Economic Association data on RCTs lists 247 RCTs conducted in India since 2012. This is the highest in the world after the US. Out of these 137 were funded by JPAL. Researchers prefer India to other poor countries. This may be because English is commonly spoken and written in most of the country, there is a fair degree of peace and safety relative to other developing countries where violence and regular work disruptions can be a problem, upsetting experiments that run for even a few short months due to political uncertainty. A study in India is cheaper than doing a study in a rich country. The cost of the programme, hiring evaluators for survey, as well as incentives for subjects can be very small sums to get participation.

As a consequence researchers come to India to undertake their research experiments. This is mainly supply driven rather than demand driven. The studies are not emerging from the problems that policy-makers are facing today. The studies are emerging from the needs of researchers to write papers for their PhDs and for publications. The bias is towards new interventions rather than trying to evaluate the impact of the same intervention in different regions, communities, and governance settings.

When foreign academics from top universities with thousands of dollars tell bureaucrats in local governments that they can put good money into a programme that will make the involved officials look good and solve some of the issues they face, the combination is often difficult to resist for the bureaucracy. Local authorities often "support" these studies by being hosts, helping make contacts or turning a blind eye to the experiment. Often Indian researchers have gravitated to becoming administrators of RCTs for foreign researchers. They are the persons on the ground managing the project while the international researcher brings funds and is the project lead. Indian researchers are sometimes co-authors, but usually only staff.

What about ethical considerations?

Furthermore, India has not put into place a law or regulations regarding how RCTs should be conducted and be kept in check. Experiments that are not possible to conduct in the US (e.g. involving minors, geotagging of individuals, the absence of informed consent) are done in India. For India (different from e.g. many other poor countries) in the absence of a regulator or an institution that gives permissions for the experiment, the procedure typically requires the university where the researcher is affiliated to approve, but no one in India sees anything going on. The visitors who might be Ph.D. students at times, come often on tourist visas as individuals without registering as a researcher, or without explicit permission for the work they are conducting. Thus, there is no mechanism for oversight of the

experiments being conducted in India/on Indian citizens. Often, consent is not taken from the subject, this violation is often not even recorded, and if recorded the subject is usually a poor person who is willing to give up privacy/information/access to their person or business for a meagre sum that is cheap for especially internally funded projects.

Due to the interests of government and academia and funding agencies, it is not in any group's interest to protect the subjects from invasive experiments by researchers. Thus, there are some kinds of things done that could be borderline unethical: feeding small children tablets, geotagging students, giving unscientific advice to farmers.

The Indian government enacted a law for clinical trials in 2017 to prevent unethical practices in clinical trials that were happening in India. Indian parliament needs to enact a law and create a regulator and a framework for social science RCTs as well.

How useful are RCTs for policy makers in India?

RCTs in development economics demonstrate what we in India might call "jugaad" economics—how to get something out of the system without actually solving the fundamental problems of economic growth or state capacity. "Jugaad" refers to finding a way to solve a problem at hand without addressing issues about why the problem arose in the first place. In general "jugaad" is not a universal solution. It is a solution that somehow manages to find a working solution to a problem. It is a work around. It is not based on addressing fundamental design issues or asking why the problem arose in the first place. It is a solution that has to be implemented again and again because the root of the problem is rarely addressed. It is just something that works for that problem at that place and time.

RCTs with their known low external validity do not provide much value for making centralized policy decisions. It is well understood that the causal claims of an RCT lack external validity, but all too often, writers and readers slip into broader claims of the world. This is a particularly important problem in India. India is a continental economy, with extreme diversity within the country. One way of thinking about this is how the ratio of the best to the worst parts of India is comparable to this ratio computed for the merger between Latin America and Africa. Because the 'country' is a natural unit of organization in the mind, we exercise caution when applying results from a paper in (say) Tanzania into a policy-making setting in (say) Chile. But we are more likely to think that a paper done in (say) Tanzania should influence all thinking in that country. As a consequence, many researchers and policy-makers tend to casually extrapolate the results of an RCT in one part of India to other parts of India. This generally leads to trouble. Writers and readers need to be more careful in how papers are written, in order to narrowly circumscribe the implications of a given research project.

In each country, at each point in time, there are certain questions that loom large in thinking about policy. As an example, in India of the recent years, the dominant problems have been those connected with bank regulation, inflation, exchange rate policy, economic regulation, and the legal system. In most of these areas, there is limited usefulness of RCTs. RCTs are useful for the officials inside a well structured regulator. As an example, the US Securities and Exchange Commission runs experiments when thinking about how to change regulations. But in India, our problems are more basic. We do not have regulatory institutions which are accountable, are enveloped in the rule of law, have checks and balances that reshape the incentives of individuals within the organization. The cutting edge of economic policy thinking lies in the administrative law frameworks to rein in the officials who wield coercive power. In this environment, there is a limited role for RCTs in influencing the thinking on the questions that matter. The disproportionate allocation of human resources in the economics community, in favor of RCTs, has helped reduce the extent to which economists have been useful in the policy process.

What is the impact of RCTs on the economics profession in India?

One of the most important sources of economic development is a capable economics community within the country. I am a bit disenchanted about the impact of the RCT revolution upon this fledgling community. As mentioned above, the disproportionate focus upon the problems that can be solved using RCTs has come at the expense of a research community that is able to work on the questions that matter. RCT papers require very large fund raising, by Indian standards. In the main, this fund raising takes place from sources outside the country. This has created a peculiar kind of dependency, where researchers who have been told that only RCT papers can publish are forced to look for the human networks which are able to raise the large volume of resources that is required for such work. This has hampered authenticity of thinking, and an emphasis on looking at the world and asking important questions. Instead of having only one distortion (the interests of editors and referees far away), there are now three additional distortions (only questions that can be answered using RCTs are eligible, only questions of interest to funders outside India are feasible, and only questions where one can find those international networks are feasible). RCT papers tend to ask very narrow questions (e.g. does teacher attendance improve when the teacher submits a selfie every morning) and this has come at the cost of broad intellectualization of the younger development economics community. There has been a decline in intellectual capacity in thinking about economics and India, in return for high career rewards for publishing in prominent journals.

Is doing RCTs and efficient allocation of scare resources for the Indian economy?

If I were a central planner allocating resources to doing economics better in India, it seems obvious to me that the highest bang for the buck lies in creating more and better data of many kinds. For example in India, we don't have good GDP, consumption, or employment data. There is a debate about how many people in India are unemployed. Policies made in the absence of data is like driving in the blind. From the perspective of an economist, it would be more useful to spend resources to measure the economy, jobs, labor markets, etc. We don't have measure of informal sector. We don't have measures of innovation or skills or productivity, which are crucial to policy-making and driving growth. We are driving in the blind when it comes to the economy as a whole. While we can find from RCTs the behavior of agents in local credit markets, we struggle with measuring trends in interest rates in informal credit markets.

Economic data for households in India is limited. As an example, it is only in recent years that the first panel dataset about households has come about in India, which observes a sample of households three times a year. Such a dataset requires resourcing that is comparable to a few RCT papers, and will create an entire body of new knowledge about India. A lot of that knowledge will be descriptive, and some of it will be quasi-experimental. This expenditure seems to be a better use of resources, when compared with a few RCT papers which would spend the same amount of money and rarely give data, and the research is non-replicable. Panel datasets create replication and competition between researchers (at modest expenditure points), in a way that RCT papers do not. Panel datasets create sustained measurement of simple facts about the country (e.g. basic facts about Women's labor force participation) which RCT projects do not. Panel datasets are much more valuable for the development of an Indian policy-relevant research community. I can see how the agency problems of government and philanthropic organizations, coupled with the solution seeking that has been fashionable in recent years, has created incentives to 'discover what works', at the expense of the more complex construction of knowledge of the humanities and social science that is required for policy-making. Many years into the RCT revolution, I would argue in favor of shifting away towards more traditional approaches.

References

Abramowicz, M. and A. Szafarz (2020). "Ethics of RCTs: Should Economists Care about Equipoise?," Chapter 10, this volume.

Acemoglu, D. (2010). "Theory, General Equilibrium, and Political Economy in Development Economics," *Journal of Economic Perspectives*, 24(3): 17–32.

Adams, V. (ed.) (2016). *Metrics: What Counts in Global Health*. Durham and London: Duke University Press.

AFD (ed.) (2012). "Unease in Evaluation: What Are the Lessons to be Drawn from the Development Experience?," *Revue d'économie du développement*, 20(4). Special issue.

Ahmed, H. M., Mitchell, M., and B. Hedt (2010). "National Implementation of Integrated Management of Childhood Illness (IMCI): Policy Constraints and Strategies," *Health Policy*, 96(2): 128–33.

Aiken, A. M., Davey, C., Hargreaves, J. R., and R. J. Hayes (2015). "Re-analysis of Health and Educational Impacts of a School-based Deworming Programme in Western Kenya: A Pure Replication," *International Journal of Epidemiology*, 44(5): 1572–80.

Akrich, M., Strum, S., Callon, M., and B. Latour (2013). *Sociologie de la traduction: textes fondateurs*. Presses des Mines via OpenEdition.

Alcott, H. (2015). "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 130(3): 1117–65.

Alfonsi, L., Bandiera, O., Bassi, V., Burgess, R., Rasul, I., Sulaiman, M., and A. Vitali (2017). "Tackling Youth Unemployment: Evidence from a Labor Market Experiment in Uganda." Working Paper, Private Enterprise Development in Low Income Countries (PEDL) Programme, DfID.

Alik Lagrange, A. and M. Ravallion (2019). "Estimating Within-Group Spillover Effects Using a Cluster Randomization: Knowledge Diffusion in Rural India," *Journal of Applied Econometrics*, 34: 110–28.

Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A., Gemmill, A., Fat, D. M., and T. Boerma (2016). "Global, Regional, and National Levels and Trends in Maternal Mortality between 1990 and 2015, with Scenario-based Projections to 2030: A Systematic Analysis by the UN Maternal Mortality Estimation Inter-Agency Group," *The Lancet*, 387(10017): 462–74.

Alkin, M. C. (2004). *Evaluation Roots: Tracing Theorists' Views and Influences*. Beverly Hills, CA: Sage Publications.

Anderson, D. M., Charles, K. K., and D. I. Rees (2018). "Public Health Efforts and the Decline in Urban Mortality." NBER Working Paper 25027.

Andrés, L., Briceño, B., Chase, C., and J. A. Echenique (2017). "Sanitation and Externalities: Evidence from Early Childhood Health in Rural India," *Journal of Water, Sanitation and Hygiene for Development*, 7(2): 272–89.

Andrews, M., Pritchett, L., and M. Woolcock (2012). "Escaping Capability Traps through Problem-Driven Iterative Adaptation." Center for Global Development Working Paper 299.

Andrews, M., Pritchett, L., and M. Woolcock (2017). *Building State Capability: Evidence, Analysis, Action*. Oxford: Oxford University Press.

Angell, M. (1997). "Editorial: The Ethics of Clinical Research in the Third World," *New England Journal of Medicine,* 337(12): 847–9.

Angelucci, M., Karlan, D., and J. Zinman (2015). "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, 7(1): 151–82.

Angrist, J. and A. Krueger (1999). "Empirical Strategies in Labor Economics," in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3. Amsterdam: North-Holland.

Angrist, J. D. and J.-S. Pischke (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.

Angrist, J., Imbens, G., and D. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, XCI: 444–55.

Angulo Salazar, L. (2013). "The Social Costs of Microfinance and Over-indebtedness for Women," in I. Guérin, S. Morvant-Roux, and M. Villarreal (eds.), *Microfinance, Debt and Over-indebtedness. Juggling with Money*. London: Routledge, 232–52.

Ansah, E. K., Narh-Bana, S., Asiamah, S., Dzordzordzi, V., Biantey, K., Dickson, K., Gyapong, J. O., Koram, K. A., Greenwood, B. M., Mills, A., and C. J. M. Whitty (2009). "Effect of Removing Direct Payment for Health Care on Utilisation and Health Outcomes in Ghanaian Children: A Randomised Controlled Trial," *PLoS Medicine*, 6(1): 0048–58.

Ardilly, P. and Tillé, Y. (2006). Sampling Methods: Exercises and Solutions (English Edition). Basingstoke: Springer.

Aristotle. *Rhétorique*. Online, http://remacle.org/bloodwolf/philosophes/Aristote/rheto1.htm.

Armendàriz, B. and J. Morduch (2010). *The Economics of Microfinance*, Second edition. Cambridge, MA: MIT Press.

Arnold, B. F., Hogan, D. R., Colford, J. M., and A. E. Hubbard (2011). "Simulation Methods to Estimate Design Power: An Overview for Applied Research," *BMC Medical Research Methodology*, 11(1): 94.

Arnold, B. F., Null, C., Luby, Stephen P., and J. M. Colford Jr. (2018). "Implications of WASH Benefits Trials for Water and Sanitation—Authors' Reply," *The Lancet Global Health*, 6(6): e616–e617.

Arnott, R. and J. E. Stiglitz (1988). "Randomization with Asymmetric Information," *RAND Journal of Economics*, 19(3): 344–62.

Arshad, A., Salam, R. A., Lassi, Z. S., Das, J. K., Naqvi, I., and Z. A. Bhutta (2014). "Community Based Interventions for the Prevention and Control of Tuberculosis," *Infectious Diseases of Poverty*, 3(1): 1–10. doi: 10.1186/2049-9957-3-27.

Arunachalam, R. S. (2011). *The Journey of Indian Micro-finance: Lessons for the Future*. Chennai: Aapti Publications.

Ashenfelter, O. C. (1983). "Determining Participation in Income-Tested Social Programs," *Journal of the American Statistical Association*, 78(383): 517–25.

Ashenfelter, O. C. and D. Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67(4): 648–60.

Athey, S. and G. W. Imbens (2017). "The Econometrics of Randomized Experiments," in A. Banerjee and E. Duflo. *The Handbook of Economic Field Experiments.* Amsterdam: North-Holland, Chapter 3: 73–140.

Athey, S. and G. W. Imbens (2018). "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." NBER Working Paper 24963.

Athey, S. and G. W. Imbens (2019). "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11(1): 685–725.

Athey, S., Chetty, R., Imbens, G., and K. Hyunseung (2016). "Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index." Unpublished manuscript.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., and H. Harmgart (2015). "The Impacts of Microfinance: Evidence from Joint-liability Lending in Mongolia," *American Economic Journal: Applied Economics*, 7(1): 90–122.

Atun, R. A., Bennett, S., and A. Duran (2008). "When Do Vertical (Stand-alone) Programmes Have a Place in Health Systems?" Paper presented at the *WHO European Ministerial Conference on Health Systems*, 1–28.

Augsburg, B., De Haas, R., Harmgart, H., and C. Meghir (2015). "The Impacts of Microcredit: Evidence from Bosnia and Herzegovina," *American Economic Journal: Applied Economics*, 7(1): 183–203.

Baele, S. (2013). "The Ethics of New Development Economics: Is the Experimental Approach to Development Economics Morally Wrong?," *Journal of Philosophical Economics*, 7(1): 1–42.

Bahadur, R. R. and L. J. Savage (1956). "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27(4): 1115–22.

Baird, S., Bohren, A., McIntosh, C., and B. Özler (2017). "Optimal Design of Experiments in the Presence of Interference," *Review of Economics and Statistics*, 100(5): 844–60.

Baldassarri, D. and M. Abascal (2017). "Field Experiments across the Social Sciences," *Annual Review of Sociology*, 43: 41–73.

Bamberger, M., Rao, V., and M. Woolcock (2010). "Using Mixed Methods in Monitoring and Evaluation, Experiences from International Development." World Bank Policy Research Working Paper 5245.

Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., and M. Sulaiman (2017). "Labor Markets and Poverty in Village Economies," *Quarterly Journal of Economics*, 132 (2): 811–70.

Banerjee, A. (2006). "Making Aid Work. How to Fight Global Poverty—Effectively," *Boston Review*, July/August.

Banerjee, A. (ed.) (2007). *Making Aid Work*. Cambridge (Massachusetts)/London: MIT Press.

Banerjee, A. (2013). "The J-PAL Story: A Decade of Partnerships," online: https://www.youtube.com/watch?v=AkC9tVUptM4&list=PL5Dr5MK6NSso3iEqn6BDu8OzyMFyLwiNE&index=5.

Banerjee, A. and E. Duflo (2009). "The Experimental Approach to Development Economics," *Annual Review of Economics*, 1: 151–78.

Banerjee, A. and E. Duflo (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, New York: Public Affairs.

Banerjee, A. and E. Duflo (2014). "The Experimental Approach to Development Economics," in D. L. Teele (ed.),*Field Experiments and Their Critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven and London: Yale University Press, 78–114.

Banerjee, A. and E. Duflo (2017). "Pushing Evidence-Based Policymaking for the Poor." Livemint, October 16.

Banerjee, A. and R. He (2008). "Making Aid Work," in W. Easterly (ed.), *Reinventing Foreign Aid*, Cambridge: The MIT Press, 47–92.

Banerjee, A. and S. Mullainathan (2010). "The Shape of Temptation: Implications for the Economic Lives of the Poor," NBER Working Paper No. 15973.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., and S. Khemani (2010). "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy,* 2: 1–30.

Banerjee, A., Breza, E., Duflo, E., and C. Kinnan (2019). "Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?," NBER Working Paper 26346.

Banerjee, A., Chassang, S., Monero, S., and E. Snowberg (2017b). "A Theory of Experimenters." NBER Working Papers 23867.

Banerjee, A., Chassang, S., and E. Snowberg (2017a). "Decision Theoretic Approaches to Experiment Design and External Validity," in A. Banerjee and E. Duflo (eds.), *Handbook of Economic Field Experiments, Volume 1*. Amsterdam: Elsevier.

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., and N. Singh (2012). "Can Institutions Be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police." NBER Working Papers 17912.

Banerjee, A., Duflo, E., and R. Glennerster (2008). "Putting a Band-aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System," *Journal of the European Economic Association,* 6: 487–500.

Banerjee, A., Duflo, E., Glennerster, R., and C. Kinnan (2015b). "The Miracle of Microfinance? Evidence from a Randomized Evaluation," *American Economic Journal: Applied Economics*, 7(1): 22–53.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and C. Udry (2015a). "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries," *Science*, 348(6236): 1–16.

Banerjee, A., Duflo, E., Keniston, D., and N. Singh (2017c). "One Question for the Road," *The Indian Express*, July 6.

Banerjee, A., Duflo, E., and M. Kremer (2016). "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy," Paper prepared for *The State of Economics, The State of the World* Conference, Proceedings Volume. September 11: 42.

Banerjee, A., Duflo, E., and M. Kremer (2019). "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy," in K. Basu, D. Rosenblatt, and C. P. Sepulveda (eds.), *State of Economics, State of the World*, Cambridge, MA: MIT Press, forthcoming.

Banerjee, A., Karlan, D., and J. Zinman (2015c). "Six Randomized Evaluations of Microcredit: Introduction and Further Steps," *American Economic Journal: Applied Economics* 7(1): 1–21.

Bardhan, P. (1984). *Land, Labor, and Rural Poverty: Essays in Development Economics*. New York, NY: Columbia University Press.

Barrett, C. and M. Carter (2010). "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections," *Applied Economic Perspectives and Policy*, 32(4): 515–48.

Barrett, C. and M. Carter (2014). "A Retreat from Radical Skepticism: Rebalancing Theory, Observational Data, and Randomization in Development Economics," in D. L. Teele (ed.), *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, New Haven and London: Yale University Press, 58–77.

Barrett, C. and M. Carter (2020). "Finding Our Balance? Revisiting the Randomization Revolution in Development Economics Ten Years Further On," *World Development*, 127 104789.

Bastiaensen, J. and P. Marchetti (2011). "Rural Microfinance and Agricultural Value Chains: Strategies and Perspectives of the Fondo de Desarrollo Local in Nicaragua," in

B. Armandariz and M. Labie (eds.), *The Handbook of Microfinance*, London and Singapore: World Scientific Publishing, 461–95.

Basu, K. (2013). "Shared Prosperity and the Mitigation of Poverty in Practice and in Precept," Policy Research Working Paper 6700, Washington DC.: The World Bank.

Basu, K. (2014). "Randomization, Causality and the Role of Reasoned Intuition," *Oxford Development Studies*, 42(4): 455–72.

Bateman, M. (2010). *Why Doesn't Microfinance Work? The Destructive Rise of Local Neoliberalism*. London: Zed Books.

Bates, M. A. and R. Glennerster (2017). "The Generalizability Puzzle," *Stanford Social Innovation Review*, Summer, 51–4.

Bates, M. A., Glennerster, R., Gumede, K., and E. Duflo (2012). "The Price Is Wrong," *The Journal of Field Action*, *Field Actions Science Reports*, Special Issue (4), Institut Veolia. http://factsreports.revues.org/1554.

Bauchet, J. and J. Morduch (2019). "Paying in Pieces: A Natural Experiment on Demand for Life Insurance under Different Payment Schemes," *Journal of Development Economics*, 139(C): 69–77.

Bauchet, J., Morduch, J., and Ravi, S. (2015). "Failure versus Displacement: Why an Innovative Anti-poverty Program Showed No Net Impact in South India," *Journal of Development Economics*, 116(C): p. 1–16.

Beacco, J.-C. and S. Moirand (1995). "Autour des discours de transmission des connaissances ," *Langages*, 117: 32–53.

Beaman, L., BenYishay, A., Magruder, J., and A. M. Mobarak (2018a). "Can Network-Theory Based Targeting Increase Technology Adoption?" NBER Working Paper No 24912.

Beaman, L., Karlan, D., Thuysbaert, B., and C. Udry (2018b). "Selection into Credit Markets: Evidence from Agriculture in Mali." Working paper.

Bédécarrats, F. (2012). "L'impact de la microfinance: un enjeu politique au prisme de ses controverses scientifiques," *Mondes en développement*, 2: 127–42.

Bédécarrats, F., Guérin, I., Morvant-Roux, S., and F. Roubaud (2019a). "Estimating Microcredit Impact with Low Take-up, Contamination and Inconsistent Data. A Replication Study of Crépon, Devoto, Duflo, and Pariente (American Economic Journal: Applied Economics, 2015)," *International Journal for Re-Views in Empirical Economics*, 3. https://www.iree.eu/publications/publications-in-iree/estimating-microcredit-impact-with-low-take-up-contamination-and-inconsistent-data-a-replication-study-of-crepon-devoto-duflo-and-pariente-american-economic-journal-applied-economics-2015/

Bédécarrats, F., Guérin, I., Morvant-Roux, S., and F. Roubaud (2019b). "Lies, Damned Lies, and RCT: une expérience de J-PAL sur le microcrédit rural au Maroc," DIAL Working Paper 2019–04.

Bédécarrats, F., Guérin, I., Morvant-Roux, S., and F. Roubaud (2019c), "Verifying the Internal Validity of a Flagship RCT: A Review of Crépon, Devoto, Duflo and Pariente. Rebutting the Rebuttal." DIAL Working Paper 2019-07B.

Bédécarrats, F., Guérin, I., and F. Roubaud (2013). " L'étalon-or des évaluations randomisées: du discours de la méthode à l'économie politique ," *Sociologies pratiques*, 2: 107–22.

Bédécarrats, F., Guérin, I., and F. Roubaud (2019). "All that Glitters Is Not Gold. The Political Economy of Randomized Evaluations in Development," *Development and Change*, 50 (3): 735–62.

Bédécarrats, F., Guérin, I., and F. Roubaud (2020). "Microfinance RCTs in Development: Miracle or Mirage?" Chapter 7, this volume.

Beisel, U. (2015). "Markets and Mutations: Mosquito Nets and the Politics of Disentanglement in Global Health." *Geoforum*, 66: 146–55.

Belissa, T., Bulte, E., Cecchi, F., Gangopadhyay, S., and R. Lensink (2019). "Liquidity Constraints, Informal Institutions, and the Adoption of Weather Insurance: A Randomized Controlled Trial in Ethiopia," *Journal of Development Economics*, 140: 269–78.

Ben David, D. and D. H. Papell (1998). "Slowdowns and Meltdowns: Postwar Growth Evidence from 74 Countries," *The Review of Economics and Statistics,* 80: 561–71.

Berg, A., Ostry, J. D., and J. Zettelmeyer (2012). "What Makes Growth Sustained?," *Journal of Development Economics,* 98: 149–66.

Bernard, T., Delarue, J., and J.-D. Naudet (2012). "Impact Evaluations: A Tool for Accountability? Lessons from Experience at Agence Française de Développement," *Journal of Development Effectiveness*, 4(2): 314–27.

Berndt, C. (2015). "Behavioural Economics, Experimentalism and the Marketization of Development," *Economy and Society*, 44(4): 567–91.

Bernhardt, A., Field, E., Pande, R. and N. Rigol (2017). "Household Matters: Revisiting the Returns to Capital among Female Micro-entrepreneurs." NBER Working Paper 23358.

Berriet-Solliec, M., Labarthe, P., and C. Laurent (2014). "Goals of Evaluation and Types of Evidence," *Evaluation*, 20(2): 195–213.

Bertrand, M., Djankov, S., Hanna, R., and S. Mullainathan (2007). "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption," *Quarterly Journal of Economics*, 122(4): 1639–76.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and J. Zinman (2010). "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *The Quarterly Journal of Economics*, 125(1): 263–306.

Beta-Blocker Heart Attack Trialists (BHAT) (1982). "A Randomized Trial of Propranolol in Patients with Acute Myocardial Infarction," *Journal of the American Medical Association*, 247(12): 1707–14.

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective.* New York: Wiley.

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and P. W. Gething (2015). "The Effect of Malaria Control on Plasmodium Falciparum in Africa between 2000 and 2015," *Nature* 526(7572): 207–11.

Bhattacharyya, O., Reeves, S., and M. Zwarenstein (2009). "What Is Implementation Research?," *Research on Social Work Practice*, 19(5): 491–502.

Bhutta, Z. A., Das, J. K., Rizvi, A., Gaffey, M. F., Walker, N., Horton, S., Webb, P., Lartey, A., and R. E. Black (2013) "Evidence-based Interventions for Improvement of Maternal and Child Nutrition: What Can Be Done and at What Cost?," *The Lancet*, 382(9890): 452–77.

Biehl, J., Petryna, A., Biehl, J., and A. Petryna. 2014. "Peopling Global Health," *Saúde e Sociedade*, 23(2): 376–89.

Bisbee, J., Dehejia, R., Pop-Eleches, C., and C. Samii (2017). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect," *Journal of Labor Economics*, 35(S1): S99–S147.

Blattman, C. (2008). Impact Evaluation 2.0, retrieved from https://www.chrisblattman.com/documents/policy/2008.ImpactEvaluation2.DFID_talk.pdf

Blustein, J. (2005). "Toward a More Public Discussion of the Ethics of Federal Social Program Evaluation," *Journal of Policy Analysis and Management*, 24(4): 824–52.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., and J. Sandefur (2018). "Experimental Evidence on Scaling Up Education Reforms in Kenya," *Journal of Public Economics*, 168: 1–20.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., Sandefur, J., DiClemente, R. J., Swartzendruber, A. L., Brown, J. L., Medeiros, M., and D. Diniz (2013). "Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education." *Sexually Transmitted Diseases*, 40(2): 111–12.

Bonds, M. H. and M. L. Rich (2018). "Integrated Health System Strengthening Can Generate Rapid Population Impacts that Can Be Replicated: Lessons from Rwanda to Madagascar," *BMJ Global Health*, 3(5), e000976.

Boone, P., Eble, A., and D. Elbourne (2013). *Risk and Evidence of Bias in Randomized Controlled Trials in Economics*. Centre for Economic Performance, LSE.

Borgerson, K. (2009). "Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine," *Perspectives in Biology and Medicine*, 52(2): 218–33.

Bornmann, L. and R. Mutz (2014). "Growth Rates of Modern Science: A Bibliometric Analysis based on the Number of Publications and Cited References," *Journal of the Association of Information Science and Technology*, 66: 2215–22.

Bothwell, L. and S. H. Podolsky (2016). "The Emergence of the Randomized, Controlled Trial," *The New England Journal of Medicine*, 375(6): 501–4.

Bothwell, L., Greene, J., Podolsky, S., and D. Jones (2016). "Assessing the Gold Standard—Lessons from the History of RCTs," *New England Journal of Medicine* 374(22): 2175–81.

Botros, S. (1990). "Equipoise, Consent and the Ethics of Randomised Clinical Trials," in P. Byrne (ed.), *Ethics and Law in Health Care and Research*, John Wiley & Sons, Chichester, UK, 9–24.

Bouguen, A., Huang, Y., Kremer, M., and E. Miguel (2019). "Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics," *Annual Review of Economics*, 11: 523–61.

Bouquet, E., Wampfler, B., Ralison, E., and M. Roesch (2007). "Trajectoires de crédit et vulnérabilité des ménages ruraux: le cas des Cecam de Madagascar," *Autrepart*, 4: 157–72.

Bourdieu, P. (1975). "The Specificity of the Scientific Field and the Social Conditions of the Progress of Reason." *Sociology of Science*, 14 (6): 19–47.

Bradford-Hill, A. (1965). "The Environment and Disease Association or Causation," *Proceedings of the Royal Society of Medicine*, 58: 295–300.

Breton, P. (1999). " La 'préférence manipulatoire' du Président du Front National ," *Mots*, 58: 101–25.

Breuer, J. B. and J. McDermott (2013). "Economic Depression in the World," *Journal of Macroeconomics*, 38: 227–42.

Broadbent, A., Vandenbroucke, J. P., and N. Pearce (2017). "Formalism or Pluralism? A Reply to Commentaries on 'Causality and Causal Inference in Epidemiology'," *International Journal of Epidemiology*, 45(6): 1841–51.

Brodeur, A., Cook, N., and A. Heyes (2018). "Methods Matter: P-Hacking and Causal Inference in Economics," IZA Working Paper 11796.

Brodeur, A., Lé, M., Sangnier, M., and Y. Zylberberg (2016). "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 8(1): 1–32.

Brody, C., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., and S. L. Dworkin (2015). "Economic Self-help Group Programs for Improving Women's Empowerment: A Systematic Review," *Campbell Systematic Reviews*, 11/1: 1–182.

Brown, C., Ravallion, M., and D. van de Walle (2018). "A Poor Means Test? Econometric Targeting in Africa," *Journal of Development Economics*, 134: 109–24.

Bruhn, M. and D. McKenzie (2009). "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics,* 1(4): 200–32.

Brummitt, C. D., Huremovic, K., Pin, P., Bonds, M. H., and F. Vega-Redondo (2017). "Contagious Disruptions and Complexity Traps in Economic Development," *Nature Human Behaviour*, 1(9): 665–72.

Bryce, J. et al. (2013). "A Common Evaluation Framework for the African Health Initiative," *BMC Health Services Research*, 13 Suppl 2(Suppl 2), p. S10.

Buera, F. J., Kaboski, J. P., and Y. Shin (2015). "Entrepreneurship and Financial Frictions: A Macrodevelopment Perspective," *Economics*, 7(1): 409–36.

Bursztyn, L., Cantoni, D., Yang, D., Yuchtman, N., and Y. J. Zhang, (2019). "Persistent Political Engagement: Social Interactions and the Dynamics of Protest Movements," NBER Conference Paper F126621.

Burtles, G. (1995). "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9(2): 63–84.

Bylander, M. (2014). "Borrowing across Borders: Migration and Microcredit in Rural Cambodia," *Development and Change*, 45(2): 284–307.

Cai, J. and A. Szeidl (2018). "Interfirm Relationships and Business Performance," *The Quarterly Journal of Economics*, 133(3): 1229–82.

Cain, G. G. (1975). "Regression and Selection Models to Improve Nonexperimental Comparisons," in C. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, New York: Academic Press: 297–317.

Cain, G. G. and H. W. Watts (1973). "Summary and Overview," in G. G. Cain. and H. W. Watts (eds.), *Income Maintenance and Labor Supply: Econometric Studies*, Chicago: Markham.

Cain, G. G. and D. A. Wissoker (1990). "A Reanalyis of Marital Stability in the Seattle-Denver Income-Maintenance Experiment," *American Journal of Sociology* 95(5): 1235–69.

Callon, M. (2006a). "Pour une sociologie des controverses technologiques ," in M. Akrich and B. Latour (ed.), *Sociologie de la traduction : Textes fondateurs*, Sciences sociales. Paris: Presses des Mines, 135–57.

Callon, M. (2006b). "Quatre modèles pour décrire la dynamique de la science," in M. Akrich and B. Latour (eds.), *Sociologie de la traduction: Textes fondateurs*, Sciences sociales. Paris: Presses des Mines, 201–51.

Camerer, C. F., Dreber, A. Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., and T. Chan (2016). "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351(6280): 1433–6.

Cameron, D. B., Mishra, A., and A. N. Brown (2016). "The Growth of Impact Evaluation for International Development: How Much Have We Learned?," *Journal of Development Effectiveness*, 8(1): 1–21.

Camfield, L. and M. Duvendack (2014). "Impact Evaluation—Are We 'Off the Gold Standard'?" *The European Journal of Development Research*, 26(1): 1–11.

Camfield, L., Duvendack, M., and R. Palmer-Jones (2014). "Things You Wanted to Know about Bias in Evaluations but Never Dared to Think," *IDS Bulletin*, 45(6): 49–64.

Campbell, D. T. (1974). *Qualitative Knowing in Action Research*, Kurt Lewin Award address, Society for the Psychological Study of Social Issues, presented at the meeting of the American Psychological Association, New Orleans, LA, September 30.

Campbell, D. T. and J. C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research,* New York: Houghton Mifflin Co., 2.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., and P. Tyrer (2000). "Framework for Design and Evaluation of Complex Interventions to Improve Health Framework for Trials of Complex Interventions," *British Medical Journal*, 321(7262): 694–6.

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., Guthrie, B., Lester, H., Wilson, P., and A. L. Kinmonth (2007) "Designing and Evaluating Complex Interventions to Improve Health Care," *British Medical Journal*, 334(7591): 455–9.

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H., McKenzie, D., and M. Mensmann (2017). "Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa," *Science*, 357(6357): 1287–90.

Caplan, A. L. (2001). "Twenty Years After: The Legacy of the Tuskegee Syphilis Study," *Bioethics, Justice and Health Care*, Belmont, CA: Wadsworth-Thomson Learning, 231–5.

Card, D. and S. DellaVigna (2013). "Nine Facts about Top Journals in Economics," NBER Working Paper 18665.

Carlson, R. V., Boyd, K. M., and D. J. Webb (2004). "The Revision of the Declaration of Helsinki: Past, Present and Future," *British Journal of Clinical Pharmacology*, 57(6): 695–713.

Cartwright, N. (2007). "Are RCTs the Gold Standard?," *BioSocieties*, 2(1): 11–20.

Cartwright, N. (2010). "What Are Randomised Controlled Trials Good For?," *Philosophical Studies*, 147(1): 59.

Cartwright, N. and J. Hardie (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford, UK: Oxford University Press.

Cartwright, N. and E. Munro (2010). "The Limitations of Randomized Controlled Trials in Predicting Effectiveness," *Journal of Evaluation in Clinical Practice*, 16(2): 260–6.

Caruso, B. A., Sclar, G., Nagel, C., Majori, F., Sola, E., Joehne, W., Deshay, R., Udaipuria, S., Williams, R., and T. Clasen (2019). *Impacts of A Multi-level Intervention, Sundara Grama, on Latrine Use and Safe Disposal of Child Faeces in Rural Odisha, India*, 3rd Grantee Final Report. New Delhi: International Initiative for Impact Evaluation.

Casaburi, L. and J. Willis (2018). Time versus State in Insurance: Experimental Evidence from Contract Farming in Kenya. *American Economic Review,* 108(12): 3778–3813.

Case, A. and A. Deaton (2015). "Rising Mortality and Morbidity among Midlife White Non-Hispanics in 21st century America," *Proceedings of the National Academy of Sciences of the USA*, 112(49): 15078–83.

Case, A. and C. Paxson (2008). "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy*, 116(3): 499–532.

Casey, K., Glennerster, R., Miguel, E., and M. Voors (2018). "Skills versus Voice in Local Development." Unpublished manuscript.

Cederlöf, G. (1997). *Bonds Lost: Subordination, Conflict and Mobilisation in Rural South India c. 1900–1970*, New-Delhi: Manohar.

Centre for Global Development (2006). *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Report of the Evaluation Gap Working Group, Washington DC, May.

Chabé-Ferret, S. (2018). "An Approach Combining Theory, Simulations and Empirics Provides Evidence of Regularities in the Bias of Observational Methods." Toulouse School of Economics.

Chakravarty, E. F. and J. F. Fries (2006). "Science as Experiment; Science as Observation," *Nature Clinical Practice Rheumatology* 2(6): 286.

Chassang, S., Miquel, P. I., and E. Snowberg (2012). "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 102(4): 1279–1309.

Chatterjee, P. (2008). "Clinical Trials in India: Ethical Concerns," *Bulletin of the World Health Organization*, 86(8): 581–2.

Chauhan, K. et al. (2019). "The 5 Star Toilet Campaign: Improving Toilet Use in Gujarat, 3rd Grantee Final Report. New Delhi: International Initiative for Impact Evaluation.

Chayanov, A. V. (1966 [1925]). *The Theory of Peasant Economy*, Homewood, IL: Richard Irwin for the American Economic Association.

Chen, S., Mu, R., and M. Ravallion (2009). "Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China," *Journal of Public Economics*, 93: 512–28.

Chenery, H., Ahluwalia, M., Bell, C., Duloy, J., and R. Jolly (1979). *Redistribution with Growth*, New York, NY: Oxford University Press.

Chernozhukov, V., Demirer, M., Duflo, E., and I. Fernandez-Val (2018). *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments*. National Bureau of Economic Research.

Cherrier, B. (2019). "Weekly Lecture Was on 'What Should Come First: Theory or Data?' So Here's Tweetstorm on the History of Quantitative Economics." *Twitter*, March 13, twitter.com/Undercoverhist/status/1105851715461570560.

Chetty, R., Hendren, N., Jones, M. R., and S. R. Porter (2019). "Race and Economic Opportunity in the United States: An Intergenerational Perspective," NBER Working Paper 24441.

Childress, J. F., Faden, R. R., Gaare, R. D., Gostin, L. O., Kahn, J., Bonnie, R. J., and P. Nieburg (2002). "Public Health Ethics: Mapping the Terrain," *Journal of Law, Medicine & Ethics*, 30(2): 170–8.

Chong, A., La Porta, R., Lopez-de-Silanes, F., and A. Shleifer (2014). "Letter Grading Government Efficiency," *Journal of the European Economics Association*, 12(2): 277–98.

Chouliaraki, L. and N. Fairclough (2010). "Critical Discourse Analysis in Organizational Studies: Towards an Integrationist Methodology." *Journal of Management Studies* 47 (6): 1213–18.

Christensen, G. and E. Miguel (2018). "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 56(3): 920–80.

Christiano, L., Eichenbaum, M., and S. Rebelo (2011). "When Is the Government Spending Multiplier Large?," *Journal of Political Economy*, 119: 78–121.

Clasen, T., Boisson, S., Routray, P., Torondel, B., Bell, M., Cumming, O., Ensink, J., Freeman, M., Jenkins, M., and M. Odagiri (2014). "Effectiveness of a Rural Sanitation Programme on Diarrhoea, Soil-Transmitted Helminth Infection, and Child Malnutrition in Odisha, India: A Cluster-Randomised Trial," *The Lancet Global Health*, 2(11): e645–e653.

Cling, J.-P., Lagrée, S., Razafindrakoto, M., and F. Roubaud (2014). *The Informal Economy in Developing Countries*. London and New York: Routledge.

Cling, J.-P., Razafindrakoto, M., and F. Roubaud (2003). *New International Poverty Reduction Strategies*. London and New York: Routledge.

Coffey, D. and D. Spears (2017). *Where India Goes: Abandoned Toilets, Stunted Development and the Costs of Caste.* London: HarperCollins.

Coffey, D. and D. Spears (2018). "Implications of WASH Benefits Trials for Water and Sanitation," *The Lancet Global Health*, 6(6): 615.

Coffey, D., Deaton, A., Drèze, J., Spears, D., and A. Tarozzi (2013). "Stunting among Children: Facts and Implications," *Economic and Political Weekly*, 48(34): 68–9.

Cohen, J. and P. Dupas (2010). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, 125(1): 1–45.

Cohen, J. and W. Easterly (2010). *What Works in Development? Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.

Collins, D., Morduch, J., Rutherford, S., and O. Ruthven (2009). *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton: Princeton University Press.

Concato, J. (2012). "Is It Time for Medicine-Based Evidence?," *The Journal of the American Medical Association*, 307(15): 1641–3.

Concato, J. (2013). "Study Design and 'Evidence' in Patient-Oriented Research," *American Journal of Respiratory and Critical Care Medicine*, 187(11): 1167–72.

Concato, J. and R. I. Horwitz (2004). "Beyond Randomised versus Observational Studies," *The Lancet*, 363(9422): 1660–1.

Concato, J. and R. I. Horwitz (2018). "Randomized Trials and Evidence in Medicine: A Commentary on Deaton and Cartwright," *Social Science & Medicine*, 210: 32–6.

Concato, J., Shah, N., and R. Horwitz (2000). "Randomized Controlled Trials, Observational Studies, and the Hierarchy of Research Design," *New England Journal of Medicine*, 342(25): 1887–92.

Congress, United States, S. C. o. F. S. o. P. A. (1978). *Welfare Research and Experimentation: Hearings Before the Subcommittee on Public Assistance of the Committee on Finance, United States Senate, Ninety-Fifth Congress, Second Session, November 15, 16, and 17*. Washington: U.S. Government.

Conlisk, J. (1973). "Choice of Response Functional Form in Designing Subsidy Experiments," *Econometrica*, 41(4): 643–56.

Conlisk, J. and H. Watts (1969). "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *American Statistical Association Proceedings, Social Statistics Section*, August, 150–6.

Cook, T. D. (2018). "Twenty-six Assumptions that Have to Be Met If Single Random Assignment Experiments Are to Warrant 'Gold Standard' Status: A Commentary on Deaton and Cartwright," *Social Science and Medicine*, 210: 37–40.

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Company.

Copestake, J., Bhalotra, S., and S. Johnson (2001). "Assessing the Impact of Microcredit: A Zambian Case Study," *The Journal of Development Studies*, 37(4): 81–100.

Copestake, J., Dawson, P., Fanning, J.-P., McKay, A., and K. Wright-Revolledo (2005). "Monitoring the Diversity of the Poverty Outreach and Impact of Microfinance: A Comparison of Methods Using Data from Peru," *Development Policy Review*, 23(6): 703–23.

Copestake, J., Johnson, S., Cabello, M., Goodwin-Groen, R., Gravesteijn, R., Humberstone, J., Nino-Zarazua, M., and M. Titus (2016). "Towards a Plural History of Microfinance," *Canadian Journal of Development Studies/Revue canadienne d'études du développement*, 37(3): 279–97.

Cornia, G., Jolly, R., and F. Stewart (eds.) (1987). *Adjustment with a Human Face: Protecting the Vulnerable and Promoting Growth*. Oxford: Oxford University Press.

Council for International Organizations and Medical Sciences (2002). *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Geneva: World Health Organization.

Coville, A. and E. Vivalt (2017). "How Often Should We Believe Positive Results? Assessing the Credibility of Research Findings in Development Economics," Unpublished manuscript.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Cox, D. R. and N. Reid (2000). *The Theory of the Design of Experiments*, Monographs on Statistics and Applied Probability 86. New York: Chapman and Hall.

Crépon, B., Devoto, F., Duflo, E., and W. Parienté (2015). "Estimating the Impact of Microcredit on Those Who Take It up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics*, 7(1): 123–50.

Crépon, B., Devoto, F., Duflo, E., and W. Parienté. (2019). "'Verifying the Internal Validity of a Flagship RCT: A Review of Crépon, Devoto, Duflo and Parienté': A Rejoinder," DIAL Working Paper 2019–07A.

Croke, K., Hicks, J. H., Hsu, E., Kremer, M., and E. Miguel (2016). "Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-Effectiveness, and Statistical Power," NBER Working Papers 22382.

Cronbach, L. (1982), *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey Bass.

Crucifix, C. and S. Morvant-Roux (2018). "Fragmented Rural Communities: The Faenas of Prospera at the Interface of Community Cooperation and State Dependency," in M. E. Balen and M. Fotta (eds.), *Money from the Government in Latin America: Conditional Cash Transfer Programs and Rural Lives*. London and New York: Routledge, 123–48.

Cull, R. and J. Morduch (2018). "Microfinance and Economic Development," in T. Beck and R. Levine (eds.), *Handbook of Finance and Development*. Cheltenham, UK: Edward Elgar.

Cull, R., Demirgüç-Kunt, A., and J. Morduch. (2018). "The Microfinance Business Model: Enduring Subsidy and Modest Profit," *The World Bank Economic Review*, 32(2): 221–44.

Cull, R., Ehrbeck, T., and N. Holle (2014). "*Financial Inclusion and Development: Recent Impact Evidence,*" CGAP, focus note 92.

Cumming, O. and V. Curtis (2018). "Implications of WASH Benefits Trials for Water and Sanitation," *The Lancet Global Health*, 6(6): e613–e614.

Cumming, O., Arnold, B. F., Ban, R., Clasen, T., Esteves Mills, J., Freeman, M. C., Gordon, B., Guiteras, R., Howard, G., and Hunter, P. R. (2019). "The Implications of Three Major New Trials for the Effect of Water, Sanitation and Hygiene on Childhood Diarrhea and Stunting—A Consensus Statement," *BMC Medicine*, 17. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1410-x

Cutler, D. and Miller, G. (2005). "The Role of Public Health Improvements in Health Advances: The Twentieth-century United States." *Demography*, 42(1): 1–22.

Czibor, E., Jimenez-Gomez, D., and J. A. List (2019). "The Dozen Things Experimental Economists Should Do (More of)," NBER Working Paper 25451.

Dahal, M. and N. Fiala (2020). *What Do We Know about the Impact of Microfinance? The Problems of Power and Precision.* World Development, 128 104773.

DARPA SCORE. "Systematizing Confidence in Open Research and Evidence (SCORE)." https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence. Last accessed: Mar. 9, 2019.

Dasgupta, P., Marglin, S., and A. Sen (1972). *Guidelines for Project Evaluation*, Vienna: United Nations Industrial Development Organization.

Datta, L.-E. (1994). "Paradigm Wars: A Basis for Peaceful Co-existence and Beyond," *New Directions for Program Evaluation*, 61: 53–70.

Davey, C., Aiken, A. M., Hayes, R. J., and J. R. Hargreaves (2015). "Re-analysis of Health and Educational Impacts of a School-Based Deworming Programme in Western Kenya: A Statistical Replication of a Cluster Quasi-Randomized Stepped-Wedge Trial," *International Journal of Epidemiology*, 44(5): 1581–92.

Davis, D. and C. A. Holt (1993). *Experimental Economics*. Princeton, NJ: Princeton University Press.

Days, S. J. and D. G. Altman (2000). "Blinding in Clinical Trials and Other Studies," *British Medical Journal*, 321: 504.

De Dickert, N. W. and E. J. Emanuel (2015). "Ethics in Cardiovascular Medicine," in D. L. Mann, D. P. Zipes, P. Libby, and R. O, Bonow (eds.), *Braunwald's Heart Disease:*

A Textbook of Cardiovascular Medicine, tenth edition, Philadelphia, PA: Elsevier Saunders, 29–34.

de Souza Leão, L. and G. Eyal (2020). "Searching under the Streetlight: A Historical Perspective on the Rise of Randomistas," *World Development*, 127 104781.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington D.C.: The World Bank.

Deaton, A. (2010a). "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48(2): 424–55.

Deaton, A. (2010b). "Understanding the Mechanisms of Economic Development," *Journal of Economic Perspectives*, 24(3): 3–16.

Deaton, A. (2012). "Searching for Answers with Randomized Experiments," Development Research Institute, NYU, video presentation https://www.youtube.com/watch?v=yiqbmiEalRU

Deaton, Angus (2013). "The Financial Crisis and the Wellbeing of Americans," *Oxford Economic Papers*, 64(1): 1–26.

Deaton, Angus (2015), "The Logic of Effective Altruism," *Boston Review*. http://bostonreview.net/forum/logic-effective-altruism/angus-deaton-response-effective-altruism

Deaton, A. and N. Cartwright (2018). "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science and Medicine*, 210: 2–21.

Deaton, A. and A. A. Stone (2016). "Understanding Context Effects for a Measure of Life Evaluation: How Responses Matter," Oxford Economic Papers, 68(4): 861–70.

Dehejia, R., Morduch, J., and H. Montgomery (2012). "Do Interest Rates Matter? Credit Demand in the Dhaka Slums," *Journal of Development Economics*, 47(2): 437–99.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). "From Local to Global: External Validity in a Fertility Natural Experiment," NBER Working Paper 21459.

DellaVigna, S. and D. Pope (2018a). "Predicting Experimental Results: Who Knows What?," *Journal of Political Economy*, 126(6): 2410–56.

DellaVigna, S. and D. Pope (2018b). "What Motivates Effort? Evidence and Expert Forecasts," *The Review of Economic Studies*, 85(2): 1029–69.

DellaVigna, S., Pope, D. and Vivalt, E. (2019). 'Predict Science to Improve Science,' *Science*, 366 (6464): 428–29.

Demirguc-Kunt, A., Klapper, L., and D. Singer (2017). *Financial Inclusion and Inclusive Growth: A Review of Recent Empirical Evidence*. Washington D.C.: The World Bank.

Desrosières, A. (1998). *The Politics Of Large Numbers: A History Of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Desrosières, A. (2013a). *Pour une sociologie historique de la quantification: l'argument statistique,* Presses des Mines via OpenEdition.

Desrosières, A. (2013b). *Gouverner par les nombres: L'argument statistique II*, Presses des Mines via OpenEdition.

Development Assistance Committee (2010). *Glossary of Key Terms in Evaluation and Results Based Management,* Paris/OECD Editions.

DFID (2012). "Broadening the Range of Designs and Methods for Impact Evaluations." Report of a Study Commissioned by the Department for International Development, DFID Working Paper 38, April.

Dhaliwal, I. and R. Hanna (2013). *Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India*. NBER Working Paper No. 20482.

Dhaliwal, I. and B. Olken (2018), *Announcing J-PAL's Policy Insights*, 1 https://www.povertyactionlab.org/blog/5-10-18/announcing-j-pals-policy-insights

Di Tillio, A., Ottaviani, M., and P. N. Sørensen (2017). "Persuasion Bias in Science: Can Economics Help?," *Economic Journal*, 127(605): F266–F304.

Dillon, A., Karlan, D., Udry, C., and J. Zinman (2020). "Good Identification, Meet Good Data," *World Development*, 127 104796.

Dimova, R. (2019). "A Debate that Fatigues…: To Randomise or Not to Randomise; What's the Real Question?," *The European Journal of Development Research*, 31(2): 163–8.

Dokova, M. (2016). "The Role of Captatio Benevolentiae in the Interaction between the Speaker and His Audience in Antiquity and Today," *Systasis* 29, Online.

Doligez, F. (2002). "Microfinance et dynamiques économiques: quels effets après dix ans d'innovations financières?," *Revue tiers monde*, 43(172) : 783–808.

Donovan, K. (2018). "The Rise of the Randomistas: On the Experimental Turn in International Aid," *Economy and Society*, 47(1): 27–58.

Doolittle, F. C. and L. Traeger (1990). *Implementing the National Jtpa Study*, New York: Manpower Demonstration Research Corporation.

Drèze, J. (2018a). "Evidence, Policy, and Politics," *Ideas for India*, August 3. https://www.ideasforindia.in/topics/miscellany/evidence-policy-and-politics.html

Drèze, J. (2018b). "Evidence, Policy and Politics: A Commentary on Deaton and Cartwright," *Social Science & Medicine*, 210: 45–7.

Dubner, S. J. (2018). Is the Protestant Work Ethic Real? Freakonomics Podcast, Episode 360. December 5, 2018. Online: http://freakonomics.com/podcast/religiosity/

Duflo, E. (2009). *Expérience, science et lutte contre la pauvreté*. Paris: Fayard.

Duflo, E. (2017). "Richard T. Ely Lecture: The Economist as Plumber," *American Economic Review: Papers and Proceedings*, 107(5): 1–26.

Duflo, E. and M. Kremer (2003). *Use of Randomization in the Evaluation of Development Effectiveness*. World Bank Operations Evaluation Department Conference on Evaluation and Development Effectiveness in Washington DC 15–16 July.

Duflo, E. and M. Kremer (2005) "Use of Randomization in the Evaluation of Development Effectiveness," in G. K. Pitman, O. N. Feinstein, and G. K. Ingram (eds.), *Evaluating Development Effectiveness*. World Bank Series on Evaluation and Development Vol. 7. New Brunswick, NJ, and London: Transaction Publishers, 205–31.

Duflo, E., Dupas, P., and M. Kremer (2015). "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," *Journal of Public Economics*, 123: 92–110.

Duflo, E., Glennerster, R., and M. Kremer (2004). "Randomized Evaluations of Interventions in Social Service Delivery," *Development Outreach*, 6(1): 26–9.

Duflo, E., Glennerster, R., and M. Kremer (2011). "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, Volume 4, Amsterdam: North-Holland.

Duflo, E., Greenstone, M., Guiteras, R., and T. Clasen (2015). "Toilets Can Work: Short and Medium Run Health Impacts of Addressing Complementarities and Externalities in Water and Sanitation," NBER Working Paper 21521.

Duflo, E., Greenstone, M., Pande, R., and N. Ryan (2013). "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *The Quarterly Journal of Economics*, 128(4): 1–49.

Duflo, E., Hanna, R., and S. P. Ryan (2012). "Incentives Work: Getting Teachers to Come to School," *American Economic Review,* 102: 1241–78.

Dumez, H. and A. Jeunemaître (2005). "La démarche narrative en économie," *Revue économique* 56(4): 983–1006.

Duncan, G., Huston, A., and T. Weisner (2007). *Higher Ground: New Hope for the Working Poor and Their Children*. New York: Russell Sage.

Dupas, P., Karlan, D., Robinson, J., and D. Ubfal (2018). "Banking the Unbanked? Evidence from Three Countries," *American Economic Journal: Applied Economics*, 10(2): 257–97.

Durbin, J. (1954). "Errors in Variables," *Review of the International Statistical Institute*, 22: 23–32.

Duteil-Mougel, C. (2005). "Les mécanismes persuasifs des textes politiques," *Corpus* 4, http://corpus.revues.org/357.

Duvendack, M. and R. Palmer-Jones (2012). "High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh," *The Journal of Development Studies*, 48(12): 1864–80.

Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., and N. Rao (2011). *What is the Evidence of the Impact of Microfinance on the Well-being of Poor People?* EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Easterly, W. (2001). *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*. Cambridge (MA): The MIT Press.

Easterly, W. (2007). *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. Oxford (UK): Oxford University Press.

Easterly, W. (2012). "If Christopher Columbus Had Been Funded by Gates," NYU Development Research Institute Blog, Online: https://nyudri.wordpress.com/2012/10/15/if-christopher-columbus-had-been-funded-by-gates/

Easterly, W. (2013). *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. New York: Basic Books.

Easterly, W. (2019). "In Search of Reforms for Growth: New Stylized Facts on Policy and Growth Outcomes." NBER Working Paper 26318.

Egil, F. (2015). "Les Objectifs de développement durable, nouveau 'palais de cristal'?" *Politique africaine*, 4: 99–120.

El-Sadr, W. M., Philip, N. M., and J. E. Justman (2014). "Letting HIV Transform Academia — Embracing Implementation Science," *New England Journal of Medicine*, 370(18): 1679–81.

Elyachar, J. (2006). *Markets of Dispossession: NGOs, Economic Development, and the State in Cairo*. Durham (NC): Duke University Press.

Elyachar, J. (2012). "Next Practices: Knowledge, Infrastructure, and Public Goods at the Bottom of the Pyramid," *Public Culture*, 24.1(66): 109–29.

Encisco, A. L. (2019). "Acaba el clientelar Prospera; surge el programa Becas Benito Juárez," *La Jornada*.

Evans, D. (2016). "That Zero Effect May Not Mean What You Think It Means, and Other Lessons from Recent Educational Research." Development Impact Blog, World Bank. January 21, 2016. Online: https://blogs.worldbank.org/impactevaluations/zero-effect-may-not-mean-what-you-think-it-means-and-other-lessons-recent-educational-research

Evans, D. and A. Popova (2016). "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews," World Bank Policy Research Working Paper 7203.

Farmer, P., Murray, M., and B. Hedt-Gauthier (2013). "Clinical Trials in Global Health Equity," *Lancet Global Health blog*. Available at: http://globalhealth.thelancet.com/2013/07/08/clinical-trials-and-global-health-equity.

Fassin, D. (2010). *La raison humanitaire. Une histoire morale du temps présent*. Paris: Gallimard/Seuil.

Faulkner, W. N. (2014). "A Critical Analysis of a Randomized Controlled Trial Evaluation in Mexico: Norm, Mistake or Exemplar?," *Evaluation*, 20(2): 230–43.

Favereau, J. (2016). "On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine," *Journal of Economic Methodology*, 23(2): 203–22.

Feinstein, A. R. and R. I. Horwitz (1997). "Problems in the 'Evidence' of 'Evidence-Based Medicine'," *The American Journal of Medicine*, 103(6): 529–35.

Ferguson, J. (1990). *The Anti-Politics Machine: "Development," Depoliticization and Bureaucratic Power in Lesotho*. Cambridge: Cambridge University Press.

Ferguson, J. (2015). *Give a Man a Fish: Reflections on the New Politics of Distribution*. Durham and Duke: Duke University Press.

Few, S. (2009). "Statistical Narrative. Telling Compelling Stories with Numbers," *Visual Business Intelligence Newsletter*, July/August: 1–10.

Field, E., Pande, R. Papp, J., and N. Rigol (2013). "Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India." *American Economic Review*, 103(6): 2196–2226.

Fiennes, C. (2018). "Funders Start Assessing Their Own Performance. To Understand What a Charity is Achieving, You Must Understand What Good Research Looks Like," *Financial Times*, November 27.

Filmer, D. and L. Pritchett (1999). "What Education Production Functions Really Show: A Positive Theory of Education Expenditures," *Economics of Education Review,* 18: 223–39.

Fine, B., Johnston, D., Santos, A. C., and E. Van Waeyenberge, E. (2016). "Nudging or Fudging: The World Development Report 2015," *Development and Change*, 47(4): 640–63.

Finkelstein, A. and S. Taubman (2015). "Randomize Evaluations to Improve Health Care Delivery," *Science*, 347(6223): 720–2.

Fisher, R. A (1926). "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33: 503–13.

Fisher, R. A. (1935). *The Design of Experiments*, London: Oliver and Boyd.

Fisher, R. A. (1960). *The Design of Experiments,* Seventh Edition, Edinburgh: Oliver and Boyd.

Fiszbein, A. and N. Schady (2010). *Conditional Cash Transfers for Attacking Present and Future Poverty,* Washington DC: World Bank.

Food and Drug Administration (2010). *Adaptive Design Clinical Trials for Drugs and Biologics*. Food and Drug Administration. Washington DC: US Government.

Ford, D. G. (2002). "Teaching Anecdotally ," *College Research*, 50(3): 114–15.

Fourcade, M., Ollion, E., and Y. Algan (2015). "The Superiority of Economists," *Journal of Economic Perspectives*, 29(1): 89–114.

Fraker, T. and R. Maynard (1987). "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, 22(2): 194–227.

Franco, A., Malhotra, N., and G. Simonovits (2014). "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science*, 345(6203): 1502–5.

Freedman, B. (1987). "Equipoise and the Ethics of Clinical Research," *The New England Journal of Medicine*, 317(3): 141–5.

Freedman, D. A. (1991). "Statistical Models and Shoe Leather," *Sociological Methodology*, 21: 291–313.

Freedman, D. H. (2010), "Lies, Damned Lies and Medical Science," *The Atlantic*, 306(4): 76–84. https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/

French, J., Blair-Stevens, C., McVey, D., and R. Merritt (2010). *Social Marketing and Public Health: Theory and Practice*. Oxford: Oxford University Press.

Frieden, T. R. (2017). "Evidence for Health Decision Making—Beyond Randomized Controlled Trials," *New England Journal of Medicine*, 377(5): 465–75.

Friedman, J. and B. Gokul (2014). "Quantifying the Hawthorne Effect," Development Impact Blog, World Bank.

Friedman, M. (2009). *Capitalism and Freedom.* Chicago: University of Chicago Press.

Friedrich, M. et al. (2019). "Promoting Latrine Use in Karnataka, India Using the RANAS Approach to Behaviour Change," 3rd Grantee Final Report. New Delhi: International Initiative for Impact Evaluation.

Fries, J. F. and E. Krishnan (2004). "Equipoise, Design Bias, and Randomized Controlled Trials: The Elusive Ethics of New Drug Development," *Arthritis Research & Therapy*, 6(3): R250.

Galasso, E. and M. Ravallion (2005). "Decentralized Targeting of an Anti-Poverty Program," *Journal of Public Economics*, 89(4): 705–27.

Galasso, E. Ravallion, M., and A. Salvia (2004). "Assisting the Transition from Workfare to Work: Argentina's *Proempleo* Experiment," *Industrial and Labor Relations Review*, 57(5): 128–42.

Garchitorena, A., Miller, A. C., Cordier, L. F., Rabeza, V. R., Randriamanambintsoa, M., Razanadrakato, H.-T. R., Hall, L., Gikic, D., Haruna, J., McCarty, M., Randrianambinina, A., Thomson, D. R., Atwood, S., Rich, M. L., Murray, M. B., Ratsirarson, J., Ouenzar, M. A., and M. H. Bonds (2018). "Early Changes in Intervention Coverage and Mortality Rates Following the Implementation of an Integrated Health System Intervention in Madagascar," *BMJ Global Health*, 3(3): e000762.

Garchitorena, A., Murray, M., Hedt-Gauthier, B., Farmer, P., and M. Bonds (2019). "Reducing the Knowledge Gap in Global Health Delivery: Contributions and Limitations of Randomized Controlled Trials," Chapter 5, this volume.

Gass, J. and L. Pritchett (2017). Returns on Scholarship (versus Organizational Learning) in Development Using (mostly) Education as an Example, *Presentation at University of Washington.*

Gautam, M. (2000). *Agricultural Extension: The Kenya Experience*, OED Precis 198, Washington DC: TheWorld Bank.

Geertz, C. (1973). *The Interpretation of Cultures.* New York: Basic Books.

Gelbach, J. B. and L. Pritchett (2002). "Is More for the Poor Less for the Poor? The Politics of Means-Tested Targeting," *The B.E. Journal of Economic Analysis & Policy,* 2(1): 1–28.

Gelman, A. (2018). "Benefits and Limitations of Randomized Controlled Trials: A Commentary on Deaton and Cartwright," *Social Science & Medicine*, 210: 48–9.

Gelman, A. and J. Carlin (2014). "Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9(6): 641–51.

Gelman, A. and F. Tuerlinckx (2000). "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures," *Computational Statistics*, 15: 373–90.

Gertler, P. (2004). "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment," *The American Economic Review*, 94(2): 336–41.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., and C. M. J. Vermeersch (2016). *Impact Evaluation in Practice,* 2nd edition, Washington, DC: Inter-American Development Bank and World Bank.

Geruso, M. and D. Spears (2018). "Neighborhood Sanitation and Infant Mortality." *American Economic Journal: Applied Economics*, 10(2): 125–62.

Ghosh, A., Gupta, A., and D. Spears (2014). "Are Children in West Bengal Shorter Than Children in Bangladesh?," *Economic & Political Weekly*, 48(8): 21–4.

Gibson, J. (2019). "Are You Estimating the Right Thing? An Editor Reflects," *Applied Economic Perspectives and Policy*, 41(3): 329–50.

Giedion, U., Alfonso, E. A., and Y. Díaz (2013). *The Impact of Universal Coverage Schemes in the Developing World: A Review of the Existing Evidence*, UNICO Studies Series (25). Washington DC.

Glennerster, R. (2012). 'The Power of Evidence: Improving the Effectiveness of Government by Investing in More Rigorous Evaluation', *National Institute Economic Review*, 219(1): R4–R14.

Glennerster, R. (2016). Not So Small. Running Randomized Evaluations, May 27, 2016, Online: http://runningres.com/blog/2016/5/27/not-so-small

Glennerster, R. and S. Powers (2016). "Balancing Risk and Benefit. Ethical Tradeoffs in Running Randomized Evaluations," in G. DeMartino and D. McCloskey (eds.), *Oxford Handbook on Professional Economic Ethics*. Oxford: Oxford University Press.

Glennerster, R. and K. Takavarasha (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton (NJ): Princeton University Press.

Glynn, A. and K. Kashin (2018). "Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program," *Journal of the American Statistical Association*, 113(523): 1040–9.

Goldberg, J. (2014). "The R-Word Is Not Dirty," Blog Post, Center for Global Development, Washington DC.

Goldberger, A. S. and C. F. Manski (1995). "Review Article: The Bell Curve by Herrnstein and Murray," *Journal of Economic Literature*, 33(2): 762–76.

Gosset, W. ("Student") (1937). "Comparison between Balanced and Random Arrangements of Field Plots," *Biometrika*, 29: 363–79.

Grasso, P. G., Wasty, S. S., and R. V. Weaving (eds.) (2003). *World Bank Operations Evaluation Department: The First 30 Years*. Washington DC: The World Bank.

Green, W. (1991). *Econometric Analysis*, New York: Macmillan.

Grosh, M. and P. Glewwe (eds.) (2000). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*, Washington DC: The World Bank.

Grossman, J. and F. Mackenzie (2005). "The Randomized Controlled Trial: Gold Standard, or Merely Standard?," *Perspectives in Biology and Medicine*, 48(4): 516–34.

Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) (1987). "Long-term Effects of Intravenous Thrombolysis in Acute Myocardial Infarction: Final Report of the GISSI Study," *Lancet,* 335(8687): 427–31.

Gubert, F. and F. Roubaud (2011). *The Impact of Microfinance Loans on Small Informal Enterprises in Madagascar: A Panel Data Analysis*. Washington DC: World Bank.

Guérin, I. and S. Kumar (2017). "Market, Freedom and the Illusions of Microcredit. Patronage, Caste, Class and Patriarchy in Rural South India," *The Journal of Development Studies*, 53(5): 741–54.

Guérin, I., Labie, M., and J.-M. Servet (eds.) (2015). *The Crises of Microcredit*, London: Zed Book.

Guérin, I., Morvant-Roux, S., and M. Villarreal (eds.) (2013). *Microfinance, Debt and Over-indebtedness: Juggling with Money*, London and New York: Routledge.

Guérin, I., Roesch, M., Venkatasubramanian, G., and S. Kumar (2013). "The Social Meaning of Over-indebtedness and Creditworthiness in the Context of Poor Rural South Indian Households (Tamil Nadu)" in I. Guérin, S. Morvant-Roux, and M. Villarreal (eds.), *Microfinance, Debt and Over-indebtedness: Juggling with Money*. London and New York: Routledge.

Guérin, I., Venkatasubramanian, G., and S. Kumar (2019). "Rethinking Saving: Indian Ceremonial Gifts as Relational and Reproductive Saving," *Journal of Cultural Economy*, forthcoming. doi.org/10.1080/17530350.2019.1583594

Gueron, J. (2017). "The Politics and Practice of Social Experiments: Seeds of a Revolution," in A. Banerjee and E. Duflo. *The Handbook of Economic Field Experiments, Vol. 1*. Amsterdam: North-Holland, Chapter 2, 27–69.

Gueron, J. and H. Rolston (2013). *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Gugerty, M. K. and D. Karlan (2018). "Ten reasons not to measure impact—And what to do instead", *Stanford Social Innovation Review*, Summer, 1–18.

Gugerty, M. K. and M. Kremer (2008). "Outside Funding and the Dynamics of Participation in Community Associations," *American Journal of Political Science*, 52(3): 585–602.

Gulhati, C. M. (2004). "Needed: Closer Scrutiny of Clinical Trials," *Indian Journal of Medical Ethics*, 1: 4–5.

Guyer, J. I. (1997). "Endowments and Assets: The Anthropology of Wealth and the Economics of Intrahousehold Allocation," in J. Haddad Lawrence, J. Hoddinott, and H. Alderman (eds.), *Intrahousehold Resource Allocation in Developing Countries* Baltimore: The Johns Hopkins University Press, 112–29.

Haavelmo, T. (1944). "The Probability Approach in Econometrics," *Econometrica*, 12(Supplement): iii–vi and 1–115.

Hahn, J., Todd, P., and W. Van der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinutiy Design," *Econometrica*, 69(1): 201–9.

Halpern, S. D., Karlawish, J. H., and J. A. Berlin (2002). "The Continuing Unethical Conduct of Underpowered Clinical Trials," *Journal of the American Medical Association*, 288(3): 358–62.

Ham, J. C. and R. J. LaLonde (1990). "Using Social Experiments to Estimate the Effect of Training on Transition Rates," in J. Hartog, G. Ridder, and J. Theeuwes (eds.), *Panel Data and Labor Market Studies*, Oxford: North-Holland, 157–72.

Hamid, S. (2019). 'Why I resigned from the Gates Foundation,' New York Times, September 26.

Hammer, J. (2014). The Chief Minister Posed Questions We Couldn't Answer. Building State Capacity Blog, Harvard University. Online: https://buildingstatecapability.com/2014/04/08/the-chief-minister-posed-questions-we-couldnt-answer/

Hammer, J. (2017). Randomized Control Trials for Development? Three Problems. https://www.brookings.edu/blog/future-development/2017/05/11/randomized-control-trials-for-development-three-problems/,' Brookings Institution Blog Post, May 11.

Hammer, J. and D. Spears (2016). "Village Sanitation and Child Health: Effects and External Validity in a Randomized Field Experiment in rural India," *Journal of Health Economics*, 48: 135–48.

Hannan, E. (2008). "Randomized Clinical Trials and Observational Studies: Guidelines for Assessing Respective Strengths and Limitations," *JACC: Cardiovascular Interventions*, 2(3): 211–17.

Hannan, M. T. and N. T. Brandon (1990). "A Reassessment of the Effect of Income Maintenance on Marital Dissolution in the Seattle-Denver Experiment," *American Journal of Sociology* 95(5): 1270–98.

Hardiman, D. (2000). *Feeding the Baniya: Peasants and Usurers in Western India*. Oxford: Oxford University Press.

Harrison, G. (2011). "Randomization and Its Discontents," *Journal of African Economies*, 20(4): 626–52. https://doi.org/10.1093/jae/ejr030.

Hathi, P., Haque, S., Pant, L., Coffey, D., and D. Spears (2017). "Place and Child Health: The Interaction of Population Density and Sanitation in Developing Countries," *Demography*, 54(1): 337–60.

Hatt, L., Chatterji, M., Miles, L., Comfort, A. B., and B. W. Bellows (2014). "A False Dichotomy: RCTs and Their Contributions to Evidence-Based Public Health," *Global Health: Science and Practice*, 3(1): 138–40.

Hatt, L., Johns, B., Connor, C., Meline, M., Kukla, M., and K. Moat (2015). *Impact of Health Systems Strengthening on Health*. Bethesda (MD): Health Finance and Governance Project, Abt Associates Inc.

Hausman, J. A. (1978). "Specification Tests in Econometrics," *Econometrica*, 46(6): 1251–72.

Hausman, J. A. and D. A. Wise (1985a). *Social Experimentation*. Chicago: University of Chicago Press.

Hausman, J. A. and D. A. Wise (1985b). "Technical Problems in Social Experimentation: Cost Versus Ease of Analysis," in J. A. Hausman and D. A. Wise (eds.). *Social Experimentation*, Chicago: University of Chicago Press, 187–220.

Hausmann, R. and C. Hidalgo (2009). "The Building Blocks of Economic Complexity." *PNAS,* 106(26): 10570–5.

Hausmann, R., Pritchett, L., and D. Rodrik (2005). Growth Accelerations. *Journal of Economic Growth,* 10(4): 303–29.

Hausmann, R., Rodrik, D., and A. Velasco (2008). "Growth Diagnostics," in N. Serra and J. Stiglitz (eds.), *The Washington Consensus Reconsidered: Towards a New Global Governance,* Oxford: Oxford University Press.

Heckman, J.J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(4): 931–59.

Heckman, J.J. (1990a). "Alternative Approaches to the Evaluation of Social Programs: Econometrics and Experimental Methods." Lecture, Sixth World Meetings of the Econometric Society, Barcelona, Spain.

Heckman, J. J. (1990b). "Varieties of Selection Bias," *American Economic Review*, 80(2: Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association): 313–18.

Heckman, J. J. (1992). "Randomization and Social Policy Evaluation," in C. F. Manski. and I. Garfinkel (eds.), *Evaluating Welfare and Training Programs*. Cambridge (MA): Harvard University Press, 201–30.

Heckman, J. J. and J. A. Smith (1995). "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2): 85–110.

Heckman, J. J. (2008). "Econometric Causality," *International Statistical Review*, 76(1): 1–27.

Heckman, J. J. (2020). "Epilogue: Randomization and Social Policy Evaluation Revisited," Chapter 12, this volume.

Heckman, J. J. and O. Ashenfelter (1973). "Estimating Labor Supply Functions," in G. G. Cain and H. Watts (eds.), *Income Maintenance and Labor Supply: Econometric Studies*. Chicago: Markham.

Heckman, J. J. and B. E. Honoré (1990). "The Empirical Content of the Roy Model," *Econometrica*, 58(5): 1121–49.

Heckman, J. J. and J. V. Hotz. (1989). "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84(408): 862–74.

Heckman, J. J. and S. Moktan (2018). "Publishing and Promotion in Economics: The Tyranny of the Top Five," NBER Working Paper 25093.

Heckman, J. J. and R. Pinto (2015). "Causal Analysis after Haavelmo," *Econometric Theory*, 31(1): 115–51.

Heckman, J. J. and R. Pinto (2019). "Exploiting Noncompliance to Enhance Causal Inference of Randomized Controlled Trials." Working paper.

Heckman, J. J. and R. Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions," in J. J. Heckman and B. S. Singer, *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press, 10: 156–245.

Heckman, J. J. and R. Robb (1986). "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in H. Wainer, *Drawing Inferences from Self-Selected Samples*, New York: Springer-Verlag, 63–107.

Heckman, J. J. and J. Smith (1995). "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2): 85–110.

Heckman, J. J. and J. Smith (1998). "Evaluating the Welfare State," in S. Strom, *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*. New York: Cambridge University Press, 241–318.

Heckman, J. J. and S. Urzúa (2010). "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156: 27–37.

Heckman, J. J. and E. Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3): 669–738.

Heckman, J. J. and E. Vytlacil (2007). "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics Volume 6B*, Amsterdam: Elsevier, 4779–4874.

Heckman, J. J., Hohmann, N., Smith, J., and M. Khoo (2000). "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, 115(2): 651–94.

Heckman, J. J., Ichimura, H., Smith, J., and P. E. Todd (1998). "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5): 1017–98.

Heckman, J. J., Ichimura, H., and P. E. Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64(4): 605–54.

Heckman, J. J., LaLonde, R. J., and J. A. Smith (1999). "The Economics and Econometrics of Active Labor Market Programs" in O. C. Ashenfelter and C. David (eds.), *Handbook of Labor Economics*, New York: North-Holland, 1865–2097.

Heckman, J. J., Lochner, L. J., and C. Taber (1998). "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, 1(1): 1–58.

Heckman, J. J., Smith, J. A., and N. Clements (1997). "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4): 487–535.

Heckman, J. J., Urzua, S., and E. Vytlacil (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3): 389–432.

Hedt-Gauthier, B. L., Chilengi, R., Jackson, E., Michel, C., Napua, M., Odhiambo, J. and A. Bawah (2017). "Research Capacity Building Integrated into PHIT Projects: Leveraging Research and Research Funding to Build National Capacity," *BMC Health Services Research*, 17(Suppl 3): 17–28.

Hemkens, L. G., Contopoulos-Ioannidis, D. G., and J. P. A. Ioannidis (2016). "Agreement of Treatment Effects for Mortality from Routinely Collected Data and Subsequent Randomized Trials: Meta-epidemiological Survey," *British Medical Journal*, 352: i493.

Hernán, M. A. and J. M. Robins (2018). *Causal Inference*. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/ Boca Raton: Chapman & Hall/CRC, forthcoming.

Hes, T. and A. Poledňáková (2013). "Correction of the Claim for Microfinance Market of 1.5 Billion Clients," *International Letters of Social and Humanistic Sciences*, 2(1): 18–31.

Hibou, B. (2011). *Anatomie politique de la domination*. Paris: La Découverte.

Hidalgo, C. A. and Hausmann, R. (2009). "The Building Blocks of Economic Complexity," *PNAS*, 106(26): 10570–5.

Hinkelmann, K. and O. Kemthorne (2008). *Design and Analysis of Experiments.* New York: John Wiley.

Hoffmann, N. (2020). "Involuntary Experiments in Former Colonies: The Case for a Moratorium," *World Development*, 127 104805.

Holland, P. W. (1986). "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81(396): 945–60.

Hopewell, S., Dutton, S., Yu, L.-M., Chan, A.-W., and D. G. Altman (2010). "The Quality of Reports of Randomised Trials in 2000 and 2006: Comparative Study of Articles Indexed in PubMed," *British Medical Journal*, 340: c723.

Horwitz, R. I. (1996). "The Dark Side of Evidence-Based Medicine," *Cleveland Clinic Journal of Medicine*, 63(6): 320–3.

Horwitz, R. I. and B. H. Singer (2017). "Why Evidence-Based Medicine Failed in Patient Care and Medicine-Based Evidence Will Succeed," *Journal of Clinical Epidemiology*, 84: 14–17.

Hotz, V. J. (1992). "Designing an Evaluation of the Job Training Partnership Act," in C. Manski and I. Garfinkel (eds.). *Evaluating Welfare and Training Programs*. Cambridge (MA): Harvard University Press, 76–114.

Houde, J.-F., Johnson, T., Lipscomb, M., and L. Schechter (2017). "Pricing Winners: Optimizing Just-in-time Procurement Auctions in Dakar, Senegal," mimeo, University of Virginia.

House, E. R. (2008). "Blowback: Consequences of Evaluation for Evaluation," *American Journal for Evaluation*, 29(4): 416–26.

House, E. R. (2014). "Origins of the Ideas in Evaluating with Validity," in J. C. Griffith and B. Montrose-Loorhead (eds.), *Revisiting Truth, Beauty and Justice: Evaluating with Validity in the 21st Century, New Directions for Evaluation*, San Francisco (CA): Jossey Bass, 142(9–10).

Hummel, A. (2013). "The Commercialization of Microcredits and Local Consumerism: Examples of Over-indebtedness from Indigenous Mexico," in I. Guérin, S. Morvant-Roux, and M. Villarreal (eds.), *Microfinance, Debt and Over-indebtedness. Juggling with money*. London: Routledge, 253–71.

Humphrey, J. H. (2009). "Child Undernutrition, Tropical Enteropathy, Toilets, and Handwashing," *The Lancet*, 374(9694): 1032–5.

Humphrey, J. H., Mbuya, M. N. N., Ntozini, R., Moulton, L. H., Stoltzfus, R. J., Tavengwa, N. V., Mutasa, K., Majo, F., Mutasa, B., and M. Goldberg(2019). "Independent and Combined Effects of Improved Water, Sanitation, and Hygiene, and Improved Complementary Feeding, on Child Stunting and Anaemia in Rural Zimbabwe: A Cluster-Randomised Trial," *The Lancet Global Health*, 7(1): e132–e147.

Humphreys, M. (2015) "What Has Been Learned from the Deworming Replications: A Nonpartisan View." Online review of data and arguments of the deworming

controversy, New York (August 18). www.columbia.edu/~mh2245/w/worms.html (accessed January 19, 2017).
IFC (2017). Strategy and Business Outlook FY18-FY20. Creating Markets and Mobilising Private Capital. Washington, DC: International Finance Corporation.
Imbens, G. (2010). "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48(2): 399–423.
Imbens, G. (2018). "Comments on Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright," *Social Science and Medicine*, 210: 50–2.
Imbens, G. and J. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2): 467–75.
Institute for Health Metrics and Evaluation (IHME) (2016). *Financing Global Health 2016. Development Assistance, Public and Private Health Spending for the Pursuit of Universal Health Coverage*. Seattle (WA): IHME.
Ioannidis, J. (2005a). "Why Most Published Research Findings are False," *PLoS Medicine*, 2(8): 1–6.
Ioannidis, J. (2005b), "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of American Medical Association*, 294(2): 218–28.
Ioannidis, J. (2018). "Randomized Controlled Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright," *Social Science & Medicine*, 210: 53–6.
Ioannidis, J., Haidich, A.-B., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. G., Contopoulos-Ioannidis, D. G., and J. Lau (2001). "Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies," *The Journal of the American Medical Association*, 286(7): 821–30.
Irwin, D. (2019). "Does Trade Reform Promote Economic Growth? A Review of Recent Evidence." NBER Working Paper 25927.
Jallais, S. (2018). "D'un monde à l'autre *ou* les rhétoriques de l'exemple dans les manuels de microéconomie," *Revue de la régulation* 23: Online.
Jamison, D., Searle, B., Galda, K., and S. P. Heyneman (1981). "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement," *Journal of Educational Psychology*, 73(4): 556–67.
Jamison, J. C. (2017). *The Entry of Randomized Assignment into the Social Sciences*. Washington DC: The World Bank.
Jatteau, A. (2013). "Expérimenter le Développement? Des économistes et leurs terrains." *Genèses* 4(93): 8–28.
Jatteau, A. (2016). "Faire preuve par le chiffre? Le cas des expérimentations aléatoires en économie," PhD Thesis, Paris Saclay.
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(165)): 94–119.
Javoy, E. and D. Rozas (2013). "Estimating Levels of Credit Market Saturation," in I. Guérin, M. Labie, and J.-M. Servet (eds.), *The Crises of Microcredit*. London: Zed Books, 39–53.
Jayachandran, S., de Laat, J., Lambin, E., Stanton, C., Audy, R., and N. Thomas (2017). "Cash for Carbon: A Randomized Trial of Payments for Ecosystem Services to Reduce Deforestation," *Science*, 357(6348): 267–73.
Jensen, R. (2010). "The (Perceived) Returns to Education and the Demand for Schooling," *Quarterly Journal of Economics*125(2): 515–48.
Jevons, W. S. (1883). *Methods of Social Reform*. London: Macmillan.

Jintarkanon, S., Nakapiew, S., Tienudom, N., Suwannawong, P., and D. Wilson (2005). "Unethical Clinical Trials in Thailand: A Community Response," *The Lancet*, 365(9471): 1617–18.

Johnson, S. and B. Rogaly (1997). *Microfinance and Poverty Reduction*. London: Oxfam.

Jones, A. and D. Steel (2018). "A Combined Theoretical and Empirical Approach to Evidence Quality Evaluation: A Commentary on Deaton and Cartwright," *Social Science and Medicine*, 210: 74–6.

Jones, B. and B. Olken (2008). "The Anatomy of Stop-Start Growth," *Review of Economics and Statistics,* 90: 582–7.

Jonston, J. (1984). *Econometric Methods,* 3rd edn. New York: McGraw Hill.

Joseph, N. (2013). "Mortgaging Used Saree-skirts, Spear-heading Resistance: Narratives from the Microfinance Repayment Standoff in Ramanagaram, India, 2008–2010," in I. Guérin, S. Morvant-Roux, and M. Villarreal (eds.), *Microfinance, Debt and Over-indebtedness. Juggling with Money*. London: Routledge, 272–94.

J-PAL (2013). "Truth-telling in Third-Party Audits." J-PAL Policy Briefcase, Cambridge (MA): Abdul Latif Jameel Poverty Action Lab.

J-PAL and IPA (2015). "Where Credit Is Due." Policy Bulletin, Cambridge (MA): Abdul Latif Jameel Poverty Action Lab and Innovations for Poverty Action.

Kabeer, N. (2019). "Randomized Control Trials and Qualitative Evaluations of a Multifaceted Programme for Women in Extreme Poverty: Empirical Findings and Methodological Reflections," *Journal of Human Development and Capabilities*, 20(2): 197–217.

Kaffenberger, M. (2018). Considering Construct Validity: Seemingly Minor Design Changes within the "Same" Project in Uganda Mait it Either the Best or Worst of all Global Literacy Interventions. *RISE,* Online: https://riseprogramme.org/blog/considering_construct_validity.

Kant, I. (1998[1785]). *Groundwork of the Metaphysics of Morals.* Cambridge, UK: Cambridge University Press.

Kaplan, R. and V. Irvin (2015). "Likelihood of Null Effects in Large NHLBI Clinical Trials Has Increased over Time," *PLoS One*, 210(8): e0132382.

Kappagoda, S. and J. Ioannidis (2014). "Prevention and Control of Neglected Tropical Diseases: Overview of Randomized Trials, Systematic Reviews and Meta-analyses," *Bulletin of the World Health Organization*, 92(5): 356–366C.

Kapur, D. (2018). 'Academic Research on India in the US: For Whom does the Bell Toll?, *India in Transition*, Center for the Advanced Study of India, University of Pennsylvania, June 29. https://theprint.in/opinion/academic-research-on-india-in-the-us-for-whom-does-the-bell-toll/78520/

Karing, A. (2018). "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," Working Paper. https://drive.google.com/file/d/1Gq59ismP9V6I2pUzuLriMVC5t6y2MqX-/view

Karlan, D. and J. Appel (2011). *More than Good Intentions: How a New Economics Is Helping to Solve Global Poverty*. New York: Dutton.

Karlan, D. and J. Zinman (2009). "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts," *The Review of Financial Studies*, 23(1): 433–64.

Karlan, D. and J. Zinman (2011). "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation," *Science*, 332(6035): 1278–84.

Karlan, D. and J. Zinman (2019). "Long-Run Price Elasticities of Demand for Credit: Evidence from a Countrywide Field Experiment in Mexico," *Review of Economic Studies*, 86(4): 1704–46.

Kass, N. (2001). "An Ethics Framework for Public Health," *American Journal of Public Health*, 91(11): 1776–82.

Kasy, M. (2016). "Why Experimenters Might Not Always Want to Randomize, and What They Should Do Instead," *Political Analysis*, 24: 324–38.

Kasy, M. and A. Sautmann (2019). "Adaptive Treatment Assignment in Experiments for Policy Choice." Preliminary draft, Harvard and MIT, June 2.

Kaufmann, D., Mehrez, G., and T. Gurgur (2002). "Voice or Public Sector Management? An Empirical Investigation of Determinants of Public Sector Performance based on a Survey of Public Officials." *World Bank Research Working Paper*.

Keane, M. (2010). "Structural vs. Atheoretic Approaches to Econometrics," *Journal of Econometrics*, 156(1): 3–20.

Keating, J. (2014). 'Random Acts. What Happens when you Approach Global Poverty as a Science Experiment? http://www.slate.com/articles/business/crosspollination/2014/03/randomized_controlled_trials_do_they_work_for_economic_development.html', *Slate*, March 26.

Kelaher, M., Ng, L., Knight, K., and A. Rahadi (2016). "Equity in Global Health Research in the New Millennium: Trends in First-Authorship for Randomized Controlled Trials among Low and Middle-Income Country Researchers 1990–2013," *International Journal of Epidemiology*, 45(6): 2174–83.

Kenny, C. and L. Pritchett (2013). "Promoting Millennium Development Ideals: The Risks of Defining Development Down," *Center for Global Development, Working Paper*.

Kerwin, J. and R. L. Thornton (2018). "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." Working paper.

Khandker, S. R., Samad, H. A., and Z. H. Khan (1998). "Income and Employment Effects of Micro-credit Programmes: Village-level Evidence from Bangladesh," *The Journal of Development Studies*, 35(2): 96–124.

Kidd, S. (2019). "The Demise of Mexico's Prospera Programme: A Tragedy Foretold," *Development Pathways*. Retrieved November 5, 2019, from https://www.development-pathways.co.uk/blog/the-demise-of-mexicos-prospera-programme-a-tragedy-foretold/

Kingi, H., Vilhuber, L., Herbert, S., and F. Stanchi (2018). "The Reproducibility of Economics Research: A Case Study." Presented at the BITSS Annual Meeting, Berkeley.

Kline, P. and C. Walters (2016). "Evaluating Public Programs with Close Substitutes: The Case of Head Start," *Quarterly Journal of Economics*, 131(4): 1795–1848.

Koetsenruijter, W. (2017). "Numbers in the News: More Ethos than Logos?," in A. Nguyen (ed.), *News, Numbers and Public Opinion in a Data-Driven World.* New York: Bloomsbury.

Kohl-Arenas, E. (2016). *The Self-Help Myth: How Philanthropy Fails to Alleviate Poverty*. Berkeley: University of California Press.

Kraay, A. (2006). When Is Growth Pro-poor? *Journal of Development Economics,* 80: 198–227.

Kramer, M. S. and S. H. Shapiro (1984). "Scientific Challenges in the Application of Randomized Trials," *JAMA: The Journal of the American Medical Association*, 252(19): 2739–45.

Kremer, M. (2003). "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons," *American Economic Review (Papers and Proceedings),* 93(2): 102–6.

Krishnaratne, S., Hensen, B., Cordes, J., Enstone, J., and J. R. Hargreaves (2016). "Interventions to Strengthen the HIV Prevention Cascade: A Systematic Review of Reviews." *The Lancet HIV* 3(7):e307–e317.

Kruk, M. E., Yamey, G., Angell, S. Y., Beith, A., Cotlear, D., Guanais, F., Jacobs, L., Saxenian, H., Victora, C., and E. Goosby (2016). "Transforming Global Health by Improving the Science of Scale-Up," *PLOS Biology*, 14(3): e1002360.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. International Encyclopedia of Unified Science. Chicago: The University of Chicago Press.

Labrousse, A. (2010). "Nouvelle économie du développement et essais cliniques randomisés: une mise en perspective d'un outil de preuve et de gouvernement," *Revue de la régulation. Capitalisme, institutions, pouvoirs* 7. Online: http://regulation.revues.org/7818.

Labrousse, A. (2016). "Not by Technique Alone. Comparing Development Analysis with Elinor Ostrom and Esther Duflo," *Journal of Institutional Economics*, 12(2): 277–303.

Labrousse, A. (2017). "Learning from Randomized Controlled Experiments. The Narrative of Scientificity, Practical Complications, Historical Experience." *Books and Ideas*, Online: http://www.booksandideas.net/Learning-from-Randomized-Controlled-Experiments.html.

Lalande, A. [1902–1923]. *Vocabulaire critique et technique de philosophie*. Paris: PUF, 6e ed.

LaLonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review* 76(4): 604–20.

Lamba, S. and D. Spears (2013). "Caste, 'Cleanliness' and Cash: Effects of Caste-based Political Reservations in Rajasthan on a Sanitation Prize," *Journal of Development Studies*, 49(11): 1592–1606.

The Lancet (2004). "The World Bank is Finally Embracing Science," *The Lancet*, 364: 731–2.

Lampedusa (di), G. T. (1960 [1958]), *The Leopard*. London, UK: Collins and Harvill Press.

Laporte, C. (2015). *L'évaluation, un objet politique: le cas d'étude de l'aide au développement,* PhD thesis. Paris: Institut d'Etudes Politiques.

Latour, B. (2012). *Enquête sur les modes d'existence. Une anthropologie des Modernes*, Paris: La Découverte.

Laura and John Arnold Foundation (2018). "Request for Proposals: Randomized Controlled Trials to Evaluate Social Programs Whole Delivery Will be Funded by Government or Other Entities." Laura and John Arnold Foundation.

Lautier, B. (2004). *L'économie informelle dans le tiers-monde*, Paris: La Découverte.

Lee, J., Morduch, J., Ravindran, S., Shonchoy, A., and H. Zaman (2018). "Poverty and Migration in the Digital Age: Experimental Evidence on Mobile Banking in Bangladesh," NYU working paper.

Lee, N. R., Rothschild, M. L., and W. Smith (2011). "Social Marketing Defined." *Social Marketing Quaterly*, Online.

Leeuw, F. and J. Vaessen (2009). *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation.* Washington DC: The World Bank.

Legovini, A., Di Maro, V., and C. Piza (2015). "Impact Evaluation Helps Deliver Development Projects," Policy Research Working Paper 7157, The World Bank.

Leigh, A. (2018). *Randomistas. How Radical Researchers Changed Our World.* New Haven: Yale University Press.

Lemieux, C. (2007). "À quoi sert l'analyse des controverses?," *Mil neuf cent. Revue d'histoire intellectuelle*, 1: 191–212.

Lensink, R. (2014). "What Can We Learn from Impact Evaluations?," *European Journal of Development Research*, 26(1): 12–17.

Levine, R. (2006). "Some Recent Developments in the International Guidelines on the Ethics of Research Involving Human Subjects," *Annals of the New York Academy of Science*, 918: 170–8.

Levy, S. (2006). *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*. Washington DC: Brookings Institution.

Lewis, A. W. (1954). "Economic Development with Unlimited Supplies of Labor," *Manchester School*, 22: 139–91.

Lilford, R. J. and J. Jackson (1995). "Equipoise and the Ethics of Randomization," *Journal of the Royal Society of Medicine*, 88(10): 552–9.

List, J. and D. Lucking-Reiley (2002). "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *Journal of Political Economy*, 110(1): 215–33.

List, J. and I. Rasul (2011). "Field Experiments in Labor Economics," in *Handbook of Labor Economics*, Volume 4, Part A, 103–228.

Little, I. and J. Mirrlees (1974). *Project Appraisal and Planning for Developing Countries*. New York: Basic Books.

London, A. J. (2017). "Equipoise in Research. Integrating Ethics and Science in Human Research," *JAMA Guide to Statistics and Methods*, 317(5): 525–6.

Luby, S. P., Rahman, M., Arnold, B. F., Unicomb, L., Ashraf, S., Winch, P. J., Stewart, C. P., Begum, F., Hussain, F., and J. Benjamin-Chung (2018). "Effects of Water Quality, Sanitation, Handwashing, and Nutritional Interventions on Diarrhoea and Child Growth in Rural Bangladesh: A Cluster Randomised Controlled Trial," *The Lancet Global Health*, 6(3): e302–e315.

Lucas, R. "A Critique on Econometric Policy Evaluation." The Philips Curve and Labor Markets, Carnegie-Rochester Conference on Public Policy. No. 1. 1976.

Lucas, R. E. (1988). "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22(1): 3–42.

Lucas, R. E. and T. J. Sargent (1981). *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press.

Lundberg, S. and J. Stearns (2019). "Women in Economics: Stalled Progress," *Journal of Economic Perspectives*, 33(1): 3–22.

Lurie, P. and S. M. Wolfe (1997). "Unethical Trials of Interventions to Reduce Perinatal Transmission of the Human Immunodeficiency Virus in Developing Countries," *New England Journal of Medicine*, 337(5): 853–6.

Mahjabeen, R. (2008). "Microfinancing in Bangladesh: Impact on Households, Consumption and Welfare," *Journal of Policy Modeling*, 30(6): 1083–92.

Marschak, J. (1953). "Economic Measurements for Policy and Prediction," in W. C. Hood and T. C. Koopmans. *Studies in Econometric Method*. New Haven (CT): Yale University Press, 1–26.

Maurer, K. and J. Pytkowska (2014). *Indebtedness of Microcredit Clients in Bosnia and Herzegovina*. Frankfurt: European Fund for Southeast Europe.

McKenzie, D. (2012). "Beyond Baseline and Follow-up: The Case for More T in Experiments," *Journal of development Economics*, 99(2): 210–21.

McKenzie, D. (2013). "How Should We Understand 'Clinical Equipoise' When Doing RCTs in Development?" Development Impact Blog, World Bank. https://blogs.worldbank.org/impactevaluations/how-should-we-understand-clinical-equipoise-when-doing-rcts-development'

McKenzie, D. (2016). *Have RCTs Taken Over Development Economics?* Development Impact, World Bank, June 13 https://blogs.worldbank.org/impactevaluations/have-rcts-taken-over-development-economics

McKenzie, D. (2018). "Six Questions with Mark Rosenzweig," *Development Impact Blog*, World Bank. January 10, 2018. https://blogs.worldbank.org/impactevaluations/six-questions-mark-rosenzweig. Last accessed November 12, 2019.

McKenzie, D. (2019). "Discussant's Comments," in K. Basu, D. Rosenblatt, and C. P. Sepulveda (eds.), *State of Economics, State of the World.* Cambridge, Mass.: MIT Press, forthcoming.

MacKenzie, D. A., F. Muniesa, and L. Siu (2007). *Do Economists Make Markets? On the Performativity of Economics*. Princeton: Princeton University Press.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., Al-Shahi Salman, R. Chan, A.-W., and P. Glasziou (2014). "Biomedical Research: Increasing Value, Reducing Waste," *The Lancet*, 383(9912): 101–4.

Mäki, U. (1995). "Diagnosing McCloskey ," *Journal of Economic Literature*, 33(3): 1300–18.

Manski, C. F. and I. Garfinkel (1992). *Evaluating Welfare and Training Programs*. Cambridge, MA / London: Harvard.

Manzi, A., Mugunga, J. C., Nyirazinyoye, L., Iyer, H. S., Hedt-Gauthier, B., Hirschhorn, L. R., and J. Ntaganira (2018). "Cost-effectiveness of a Mentorship and Quality Improvement Intervention to Enhance the Quality of Antenatal Care at Rural Health Centers in Rwanda," *International Journal for Quality in Health Care*, 31(5): 359–64.

Manzi, A., Nyirazinyoye, L., Ntaganira, J., Magge, H., Bigirimana, E., Mukanzabikeshimana, L., Hirschhorn, L. R., and B. Hedt-Gauthier (2018). "Beyond Coverage: Improving the Quality of Antenatal Care Delivery through Integrated Mentorship and Quality Improvement at Health Centers in Rural Rwanda," *BMC Health Services Research*, 18(1): 1–8.

Martinez-Alonso, E. and J. M. Ramos (2016). "A Systematic Review of Randomized Clinical Trials Published in Malaria Journal between 2008 and 2013," *Rev Esp Quimioter*, 29(3): 130–45.

Maurer, K. and J. Pytkowska (2011). *Indebtedness of Microcredit Clients in Bosnia and Herzegovina*. Frankfurt: European Fund for Southeast Europe.

McCloskey, D. (1983), "The Rhetoric of Economics," *Journal of Economic Literature*, 21(2): 481–517.

McMurray, J. J. V. (2010). "Systolic Heart Failure," *New England Journal of Medicine*, 362: 228–38.

Meager, R. (2019). "Understanding the Average Impact of Microcredit Expansion: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, 11(1): 57–91.

Meessen, B., Hercot, D., Noirhomme, M., Ridde, V., Tibouti, A., Tashobya, C. K., and L. Gilson (2011). "Removing User Fees in the Health Sector: A Review of Policy Processes in Six Sub-Saharan African Countries," *Health Policy and Planning*, 26(Suppl. 2): ii16–ii29.

Meier, G. M. and D. Seers (eds.) (1984). *Pioneers in Development*. Oxford: Oxford University Press.

Meyer, M., Heck, P., Holtzman, G., Anderson, S., Cai, W., Watts, D., and C. Chabris (2019). "Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments," *PNAS*, 116 (22): 10723–8.

Miguel, E. and M. Kremer (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1): 159–217.

Miles, S. H. (2005). *The Hippocratic Oath and the Ethics of Medicine*. Oxford: Oxford University Press.

Milford, C., Wassenaar, D., and C. Slack (2006). "Resources and Needs of Research Ethics Committees in Africa: Preparations for HIV Vaccine Trials," *IRB: Ethics & Human Research*, 28(2): 1–9.

Mill, J. S. (1843). *A System of Logic, Ratiocinative and Deductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Evidence*. London: John Parker, Eight edition available at: https://www.gutenberg.org/files/27942/27942-pdf.pdf

Miller, F. G. and H. Brody (2007). "Clinical Equipoise and the Incoherence of Research Ethics," *Journal of Medicine and Philosophy*, 32(2): 151–65.

Miller, F. G. and S. Joffe (2011). "Equipoise and the Dilemma of Randomized Clinical Trials," *New England Journal of Medicine*, 364(5): 476–80.

Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K. Harris, M. W., Ehrlich Sachs, S., Stuart, E. A., and A. Feller (2018). "The Millennium Villages Project: A Retrospective, Observational, Endline Evaluation," *The Lancet Global Health*, 6(5): e500–e513. doi: 10.1016/S2214-109X(18)30065–2.

Moffitt, R. (2004). "The Role of Randomized Field Trials in Social Science Research," *American Behavioral Scientist*, 47(5): 506–40.

Moffitt, R. (2006). "Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective," in B. Schneider and S.-K. McDonald (eds.), *Scale-Up in Education: Ideas in Principle*. Lanham, Maryland: Rowman and Littlefield.

Monte, M. (2007). "L'oxymore: figure syntactico-sémantique ou élément d'une stratégie para-doxique? ," *Fabula*, https://www.fabula.org/atelier.php?L%27oxymore%3A_%26eacute%3Bl%26eacute%3Bment_d%27une_strat%26eacute%3Bgie_para%2Ddoxique%3F. Last accessed November 12, 2019.

Morgan, K. L. and D. B. Rubin (2012). "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics*, 40(2): 1263–82.

Morduch, J. (1999). "The Microfinance Promise," *Journal of Economic Literature*, 37(4): 1569–1614.

Morduch, J. (2000). "The Microfinance Schism," *World Development,* 28(4): 617–29.

Morduch, J. (2020a). "The Disruptive Power of RCTs," Chapter 3, this volume.

Morduch, J. (2020b). "Why RCTs Failed to Answer the Biggest Questions about Microcredit Impact," *World Development*, 127 104818.

Morvant-Roux, S. (ed.) (2009). *Exclusion et liens financiers: microfinance pour l'agriculture des pays du Sud*, Paris: Economica.

Morvant-Roux, S. (2013). "International Migration and Over-indebtedness in Rural Mexico," in I. Guérin, S. Morvant-Roux, and M. Villarreal (eds.), *Microfinance, Debt and Over-Indebtedness: Juggling with Money*. Routledge: London and New York, 170–92.

Morvant-Roux, S. and M. Roesch (2015). "The Social Credibility of Microcredit in Morocco after the Default Crisis," in I. Guérin, M. Labie, and J.-M Servet (eds.), *The Crises of Microcredit*: Zed Book: London, 113–30.

Morvant-Roux, S., Guérin, I., Roesch, M., and J.-Y. Moisseron (2014). "Adding Value to Randomization with Qualitative Analysis: The Case of Microcredit in Rural Morocco," *World Development*, 56: 302–12.

Mosse, D. (2004). *Cultivating Development: An Ethnography of Aid Policy and Practice*. London: Pluto Press.

Mudur, G. (2005). "India Plans to Audit Clinical Trials," *British Medical Journal*, 331(7524): 1044.

Mueller, U. (2020). "A more robust t-test," https://www.princeton.edu/~umueller/heavy-mean.pdf.

Müller O., De Allegri, M., Becher, H., Tiendrebogo, J., Beiersmann, C., Ye, M., Kouyate, B., Sie, A., and A. Jahn (2008). "Distribution Systems of Insecticide-Treated Bed Nets for Malaria Control in Rural Burkina Faso: Cluster-Randomized Controlled Trial," *PLoS ONE*, 3(9): e3182.

Mulligan, C. (2014). "The Economics of Randomized Experiments," *Economix Blog, New York Times*, March 5.

Mummolo, J. and E. Peterson (2019). "Demand Effects in Survey Experiments: An Empirical Assessment," *American Political Science Review*, 113(2): 517–29.

Muralidharan, K., Niehaus, P., and S. Sukhtankar (2016). "Building State Capacity: Evidence from Biometric Smartcards in India. *American Economic Review,*" 106(10): 2895–2929.

Muralidharan, K., Niehaus, P., and S. Sukhtankar (2018). "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India." NBER Working Paper 23838.

Muralidharan, K., Niehaus, P., Sukhtankar, S., and J. Weaver (2018a). "Improving Last-Mile Service Delivery Using Phone-Based Monitoring." NBER Working Paper 25298.

Muralidharan, K., Niehaus, P., Sukhtankar, S., and J. Weaver (2018b). "Use Mobiles to Improve Governance," *Hindustan Times*, December 5.

Murgai, R., Ravallion, M., and D. van de Walle (2015). "Is Workfare Cost Effective against Poverty in a Poor Labor-Surplus Economy?," *World Bank Economic Review*, 30(3): 413–45.

Musgrave, R. (2008). "Merit Goods," in S. N. Durlauf and L. E. Blume (eds.), *The New Palgrave Dictionary of Economics: Volume 1–8*, London: Palgrave Macmillan UK, 4173–6.

Nadel, S. and L. Pritchett (2016). "Searching for the Devil in the Details: Learning about Development Program Design," *Center for Global Development Working Papers,* 434.

National Bioethics Advisory Commission (NBAC) (2001). *Ethical and Policy Issues in Research Involving Human Participants*, Volume I. Bethesda, MD.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, Washington DC: US Department of Health, Education and Welfare.

Narita, Y. (2018). "Toward an Ethical Experiment," *Cowles Foundation Discussion Paper No. 2127*, Yale University.

Narotzky, S. and N. Besnier (2014). "Crisis, Value, and Hope: Rethinking the Economy: An Introduction to Supplement 9," *Current Anthropology*, 55(S9): S4–16.

Naudet, J.-D. (1999). *Trouver des problèmes aux solutions. Vingt ans d'aide au Sahel*. Paris: OECD Editions.

Naudet, J.-D. (2006). "Les OMD et l'aide de cinquième génération," *Afrique contemporaine*, 2: 141–74.

Niehaus, P. (2019). "RCTs: Why Scale Matters," *VoxDev video*, https://youtu.be/fD6MgGM5jWI

Nies, A. S., Evans, G. H., and D. G. Shand (1973). "Regional Hemodynamic Effects of Beta-Adrenergic Blockade with Propranolol in the Unanesthetized Primate," *American Heart Journal*, 85: 97–102.

Null, C., C. P. Stewart, A. J. Pickering, H. N. Dentz, B. F. Arnold, C. D. Arnold, J. Benjamin-Chung, T. Clasen, K. G. Dewey, and L. C. H. Fernald (2018). Effects of Water Quality, Sanitation, Handwashing, and Nutritional Interventions on Diarrhoea and Child Growth in Rural Kenya: A Cluster-Randomised Controlled Trial. *The Lancet Global Health*, 6(3): e316–e329.

Odhiambo, J., Amoroso, C. L., Barebwanuwe, P., Warugaba, C., and B. L. Hedt-Gauthier (2017). "Adapting Operational Research Training to the Rwandan Context: Intermediate Operational Research Training Programme," *Global Health Action*, 10(1). https://www.tandfonline.com/doi/full/10.1080/16549716.2017.1386930

Ogden, T. N. (2017). *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. Cambridge, Massachusetts: MIT Press.

Ogden, T. N. (2020). "RCTs in Development Economics, Their Critics and Their Evolution." Chapter 4, this volume.

Opem, L. C. and N. Goronja (2013). *Responsible Finance: Reducing Over-indebtedness for Bosnia and Herzegovina's Microfinance Borrowers*. Washington, D.C: International Finance Corporation.

Orcutt, G. H. and A. G. Orcutt (1968). "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," *American Economic Review*, 58(4): 754–72.

Orr, R., Pang, N., Pellegrino, E. and M. Siegler (1997). "Use of the Hippocratic Oath: A Review of Twentieth Century Practice and a Content Analysis of Oaths Administered in Medical Schools in the U.S. and Canada in 1993," *Journal of Clinical Ethics* 8(4): 377–88.

Özler, B. (2018). "Incorporating Participants Welfare and Ethics into RCTs." Development Impact Blog Post, World Bank.

Palca, J. (1989). "AIDS Drug Trials Enter New Age," *Science, New Series*, 246(4926): 19–21.

Pamiès-Sumner, S. (2015). "Development Impact Evaluation, State of Play and New Challenges." A Savoir. Paris: AFD.

Parfit, D. (2011). *On What Matters.* Oxford, UK: Oxford University Press.

Patil, S. R., Arnold, B. F., Salvatore, A.L., Briceno, B., Ganguly, S., Colford Jr, J. M., and P. J. Gertler (2014). "The Effect of India's Total Sanitation Campaign on Defecation Behaviors and Child Health in Rural Madhya Pradesh: A Cluster Randomized Controlled Trial," *PLoS medicine*, 11(8): e1001709.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd edition). New York: Cambridge University Press.

Pearl, J. and D. Mackenzie (2018). *The Book of Why. The New Science of Cause and Effect.* New York: Basic Books.

Peters, J., Langbein, J. and G. Roberts (2018). "Generalization in the Tropics—Development Policy, Randomized Controlled Trials, and External Validity," *World Bank Research Observer*, 33(1): 34–64.

Petryna, A. (2007). "Clinical Trials Offshored: On Private Sector Science and Public Health," *BioSocieties*, 2(1): 21–40.

Petryna, A. (2009). *When Experiments Travel: Clinical Trials and the Global Search for Human Subjects.* Princeton, NJ: Princeton University Press.

Petticrew, M., McKee, M., Lock, K., Green, J., and G. Phillips (2013). "In Search of Social Equipoise." *British Medical Journal* (Online), 347.

Philibert, A., Ravit, M., Ridde, V., Dossa, I., Bonnet, E., Bédecarrats, F., and A. Dumont (2017). "Maternal and Neonatal Health Impact of Obstetrical Risk Insurance Scheme in Mauritania: A Quasi Experimental Before-and-After Study," *Health Policy and Planning*, 32(3): 405–17.

Picciotto, J. (2011). *Labors of Innocence*. Cambridge, Massachusetts: Harvard University Press.

Picciotto, R. (2012). "Experimentalism and Development Evaluation: Will the Bubble Burst?," *Evaluation*, 18(2): 213–29.

Picciotto, R. (2013). "Evaluation Independence in Organizations," *Journal of Multi-Disciplinary Evaluation,* Western Michigan University, 9(20), February 19.

Picherit, D. (2018). "Rural Youth and Circulating Labour in South India: The Tortuous Paths towards Respect for Madigas," *Journal of Agrarian Change*, 18(1): 178–95.

Pickering, A.J. et al. (2019). "The WASH Benefits and SHINE Trials: Interpretation of WASH Intervention Effects on Linear Growth and Diarrhoea," *Lancet Global Health*, 7(8): e1139–e1146.

Pinkovskiy, M. and X. Sala-I-Martin (2016). "Lights, Camera…Income! Illuminating the National Accounts-Household Surveys Debate," *The Quarterly Journal of Economics,* 131: 579–631.

Pitt, M. M. and S. R. Khandker (1998). "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?," *Journal of Political Economy*, 106(5): 958–96.

Plsek, P. E. and T. Greenhalgh (2001). "Complexity Science: The Challenge of Complexity in Health Care," *British Medical Journal*, 323(7313): 625–8.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., and N. P. Podsakoff (2003). "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology*, 88(5): 879.

Porter, T. (1995). *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Pradhan, M., Suryahadi, A., Sumarto, S., and L. Pritchett (2001). "Eating Like Which 'Joneses?' An Iterative Solution to the Choice of a Poverty Line 'Reference Group,'" *Review of Income and Wealth*, 47(4): 473–87.

Prasad, V., Jorgenson, J., Ioannidis, J. P. A., and A. Cifu (2013). "Observational Studies Often Make Clinical Practice Recommendations: An Empirical Evaluation of Authors' Attitudes", *Journal of Clinical Epidemiology*, 66(4): 361–6.e4.

Prathap, V. and R. Khaitan (2016). "When Is Microcredit Unsuitable? Guidelines Using Primary Evidence from Low-Income Households in India," *IFF Working Paper Series No. WP-2016–01*. Chennai: IFMR Finance Foundation.

Pritchett, L. (2000). "Understanding Patterns of Economic Growth: Searching for Hills among Mountains, Plateaus, and Plains," *World Bank Economic Review,* 14: 221–50.

Pritchett, L. (2005). "The Political Economy of Targeted Safety Nets." World Bank Social Protection Discussion Paper Series 501.

Pritchett, L. (2006). "Who Is Not Poor? Dreaming of a World Truly Free of Poverty," *The World Bank Research Observer,* 21: 1–23.

Pritchett, L. (2010a) "Is Microfinance a Shumpeterian Dead End?," Center for Global Development Blog, March 10, 2010. http://blogs.cgdev.org/open_book/2010/05/is-microfinance-a-schumpeterian-dead-end.php. Last accessed August 15, 2019.

Pritchett, L. (ed.) (2010b). *The Policy Irrelevance of the Economics of Education: Is "Normative as Positive" Useless or Worse,* Washington DC: Brookings Press.

Pritchett, L. (2013a). *The Dangerous Seduction of the Kinky*. Cambridge MA: Center for International Development.

Pritchett, L. (2013b). "RCTs in Development, Lessons from the Hype Cycle," Center for Global Development Blog, November 14, 2013. https://www.cgdev.org/blog/rcts-development-lessons-hype-cycle. Last accessed November 12, 2019.

Pritchett, L. (2014a). "Is Your Impact Evaluation Asking Questions that Matter?," *Center for Global Devleopment Blog*, November 6, 2014 https://www.cgdev.org/blog/your-impact-evaluation-asking-questions-matter-four-part-smell-test. Last accessed November 12 2019.

Pritchett, L. (2014b). "A Development Agenda without Developing Countries? The Politics of Penurious Poverty Lines (Part I)," *Center for Global Development Blog*. September 4, 2014. https://www.cgdev.org/blog/development-agenda-without-developing-countries-politics-penurious-poverty-lines-part-i. Last accessed November 12, 2019.

Pritchett, L. (2014c). "An Homage to the Randomistas on the Occasion of the J-PAL 10th Anniversary: Development as a Faith-Based Activity." Center for Global Development Blog,March10,2014.https://www.cgdev.org/blog/homage-randomistas-occasion-j-pal-10th-anniversary-development-faith-based-activity. Last accessed November 12, 2019.

Pritchett, L. (2015). "Can Rich Countries Be Reliable Partners for National Development?," *Horizons: Journal of International Relations and Sustainable Development,* 2: 206–23.

Pritchett, L. (2016). "Turns out Development Does Bring Development," *Center for Global Development Blog,* September 21, 2016. https://www.cgdev.org/blog/turns-out-development-does-bring-development. Last accessed November 12, 2019.

Pritchett, L. (2017). "The Perils of Partial Attribution: Let's All Play for Team Development," *Center for Global Development Blog,* October 26, 2017. https://www.cgdev.org/publication/perils-partial-attribution. Last accessed November 12, 2019.

Pritchett, L. (2018a). "The Debate about RCTs in Development Is Over: We Won. They Lost," Lecture, New York: DRI.

Pritchett, L. (2018b). "Knowledge or Its Adoption?," *Center for Global Development Blog*, August 6, 2018. https://www.cgdev.org/publication/knowledge-or-its-adoption. Last accessed November 12, 2019.

Pritchett, L. (2018c). We Knew Fire Was Hot. *RISE Blog*. https://www.riseprogramme.org/publications/we-knew-fire-was-hot. Last access November 12, 2019.

Pritchett, L. (2020). "Randomizing Development: Method or Madness?," Chapter 2, this volume.

Pritchett, L. and J. Sandefur (2013a). "Claims to External Validity and Development Practice Don't Mix: Theory, Simulation, and Empirics," *Journal of Globalization and Development,* 4: 161–97.

Pritchett, L. and J. Sandefur (2013b). "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix," *Center for Global Development Working Paper*, 336.

Pritchett, L. and J. Sandefur (2015). "Learning from Experiments when Context Matters," *American Economic Review: Papers and Proceedings*, 105(5): 471–5.

Pritchett, L. and L. H. Summers (2014). "Asiaphoria Meets Regression to the Mean," NBER Working Papers 20573.

Pritchett, L., Samij, S., and J. Hammer (2012). "It's All about MeE: Using Structured Experiential Learning ('e') to Crawl the Design Space," *HKS Faculty Research Working Paper Series*.

Pritchett, L., Samij, S., and J. Hammer (2013). "It's All about MeE: Learning in Development Projects through Monitoring ('M'), Experiential Learning ('e') and Impact Evaluation ('E'). Center for Global Development Working Paper 233.

Pritchett, L., Sen, S., Kar, S., and Raihande S. (2016). "Trillions Gained and Lost: Estimating the Magnitudes of Growth Episodes," *Economic Modelling*, 55: 279–91.

Pulla, P. (2018). "Link between Sanitation, Stunting Questioned," *The Hindu*. 3 February.

Putnam, H. (2009). *Renewing Philosophy*, Harvard: Harvard University Press.

Quentin, A. and I. Guérin (2013). "La randomisation à l'épreuve du terrain," *Revue Tiers Monde*, 1: 179–200.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5): 1281–302.

Rai, A. and T. Sjöström (2004). "Is Grameen Lending Efficient? Repayment Incentives and Insurance in Village Economies," *Review of Economic Studies*, 71(1): 217–34.

Ramey, C. T., Collier, A. M., Sparling, J. J., Loda, F. A., Campbell, F. A., Ingram, D. A. and N. W. Finkelstein (1976). "The Carolina Abecedarian Project: A Longitudinal and Multidisciplinary Approach to the Prevention of Developmental Retardation," in T. Theodore (ed.). *Intervention Strategies for High-Risk Infants and Young Children*, Baltimore (MD): University Park Press, 629–55.

Rao, V. (2001). "Celebrations as Social Investments: Festival Expenditures, Unit Price Variation and Social Status in Rural India," *Journal of Development Studies*, 38(1): 71–97.

Ravallion, M. (2009a). "Should the Randomistas Rule?," *Economists' Voice*, 6(2): 1–5.

Ravallion, M. (2009b). "Evaluation in the Practice of Development," *World Bank Research Observer*, 24(1): 29–54.

Ravallion, M. (2012). "Fighting Poverty one Experiment at a Time: A Review Essay on Abhijit Banerjee and Esther Duflo, *Poor Economics*," *Journal of Economic Literature*, 50(1): 103–14.

Ravallion, M. (2014). "On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments," *Journal of Econometric Methods* 4(1): 145–51.

Ravallion, M. (2016). *The Economics of Poverty: History, Measurement and Policy*, New York: Oxford University Press.

Ravallion, M. (2020). "Should the Randomistas (Continue to) Rule?," Chapter 1, this volume.

Ravallion, M., van de Walle, D., Dutta, P., and R. Murgai (2015). "Empowering Poor People through Public Information? Lessons from a Movie in Rural India," *Journal of Public Economics*, 132(December): 13–22.

Ravit, M., Ravalihasy, A., Audibert, M., Ridde, V., Bonnet, E., Raffalli, B., Roy, F.-A., N'Landu, A., and A. Dumont (2020). "The Impact of the Obstetrical Risk Insurance Scheme in Mauritania on Maternal Healthcare Utilization: A Propensity Score Matching Analysis," *Health Policy and Planning*. https://doi.org/10.1093/heapol/czz150

Rawlins, M. (2008). "De Testimonio: On the Evidence for Decisions about the Use of Therapeutic Interventions," *The Lancet*, 372(9656): 2152–61.

Raynaud, D. (2018). *Sociologie des controverses scientifiques*. Paris: Éditions Matériologiques.

Reddy, S. and R. Lahoty (2016). "$1.90 a day: What Does it Say? The New International Poverty Line," *New Left Review*, 97: 106–27.

Redfield, P. (2012). "Bioexpectations: Life Technologies as Humanitarian Goods," *Public Culture* 24 (1(66)): 157–84.

Reidy, W. J., Rabkin, M., Syowai, M., Schaaf,A., and W. M. El-Sadr (2018) "Patient-level and Program-level Monitoring and Evaluation of Differentiated Service Delivery for HIV: A Pragmatic and Parsimonious Approach Is Needed," *AIDS (London, England)*, 32(3): 399–401.

Requejo, J. H., Bryce, J., Barros, A. J. D., Berman, P., Bhutta, Z., Chopra, M., Daelmans, B., De Francisco, A., Lawn, J., and B. Maliqi (2015). "Countdown to 2015 and Beyond: Fulfilling the Health Agenda for Women and Children," *The Lancet*, 385(9966): 466–76.

Revel, J. (ed.) (1996). *Jeux d'échelles. La micro-analyse à l'expérience*. Paris: Gallimard-Le Seuil.

Ridde, V. and S. Haddad (2009) "Abolishing User Fees in Africa," *PLoS Medicine*, 6(1), p. e1000008.

Rioux, R. (2019), *Réconciliations*. Paris: Débats Publics Editions.

Rodrick, D. (2008). *One Economics, Many Recipes*, *Globalization, Institutions, and Economic Growth*. Princeton: Princeton University Press.

Rodrik, D. (2009). "The New Development Economics: We Shall Experiment, but How Shall We Learn?," in J. Cohen and W. Easterly (eds.), *What Works in Development? Thinking Big and Thinking Small*, Washington, D.C.: Brookings Institution Press.

Roethlisberger, F. J. and W. J. Dickson (1939). *Management and the Worker*. Volume 5, Cambridge, Mass.: Harvard University Press.

Roodman, D. and J. Morduch (2014). "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," *Journal of Development Studies*, 50(4): 583–604.

Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70: 41–55.

Rosenberg, R. (2009). "The New Moneylenders: Are the Poor Being Exploited by High Microcredit Interest Rates?," *CGAP Occasional Paper 15*. Washington, DC: Consultative Group to Assist the Poor.

Rosenboom, J. W. and R. Ban (2017). "From New Evidence to Better Practice: Finding the Sanitation Sweet Spot," *Waterlines*, 36(4): 267–83.

Royal Academy of Sciences (2019). "The Prize in Economic Sciences 2019" Stockholm: Press Release. https://www.nobelprize.org/uploads/2019/10/press-economicsciences2019-2.pdf.

Royal Swedish Academy of Sciences. (2019). "Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019. Understanding Development and Poverty Alleviation."

Rozas, D. (2014). "Microfinance in Mexico: Beyond the Brink," *European Microfinance Platform Blog*, June 6, 2014. http://www.e-mfp.eu/blog/microfinance-mexico-beyond-brink. Last accessed November 12, 2019.

Rubin, D. B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6(1): 34–58.

Ruhm, C. J. (2019). Shackling the Identification Police?', *Southern Economic Journal*, 85(4): 1016–26.

Russell, B. (1912). *The Problems of Philosophy*, New York: Henry Holt and Co.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D., Greaves, F., Harper, L., Hawe, P., Moore, L., Petticrew, M., Rehfuess, E., Shiell, A., Thomas, J. and M. White (2017). "The Need for a Complex Systems Model of Evidence for Public Health," *The Lancet*, 390(10112): 2602–4.

Sabet, S. M. and A. Brown (2018). "Is Impact Evaluation Still on the Rise? The New Trends 2010–2015," *Journal of Development Effectiveness*, 10(3): 291–304.

Sachs, J. (2001) *Macroeconomics and Health: Investing in Health for Economic Development. Report of the Commission on Macroeconomics and Health*, Geneva, Switzerland.

Sachs, J. (2005). *The End of Poverty: Economic Possibilities for Our Time.* New York: Penguin.

Samii, C. (2016). "Causal Empiricism in Quantitative Research," *Journal of* Politics, 78(3): 941–55.

Sandefur, J. (2015). "The Final Word on Microcredit?" *Center for Global Development Blog*; January 22. https://www.cgdev.org/blog/final-word-microcredit. Last accessed November 12, 2019.

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., and C. D'Este (2007). "Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions," *American Journal of Preventive Medicine*, 33(2): 155–61.

Sardan, J.-P. Olivier de. (1995). *Anthropologie et développement: essai en socio-anthropologie du changement social*. Paris: Karthala.

Sarin, A. (2019). "Indecent Proposals in Economics," *The India Forum*, Nov 1, Online: https://www.theindiaforum.in/article/indecent-proposals-economics

Sathyamala, C. (2019). "In the Name of Science: Ethical Violations in the ECHO Randomised Trial,'"*Global Public Health*, forthcoming. doi.org/10.1080/1744169 2.2019.1634118

Savage, L. J. (1962). *The Foundations of Statistical Inference: A Discussion Opened by L.j. Savage at the Meeting of the Joint Statistics Seminar, Birkbeck and Imperial Colleges, in the University of London*, New York: Barnes and Noble.

Savedoff, W. D., Levine, R., and N. Birdsall (eds.) (2006) *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, DC: Center for Global Development.

Schafer, A. (1982). "The Ethics of the Randomized Clinical Trial," *New England Journal of Medicine* 307(12): 719–24.

Schicks, J. (2013). "The Definition and Causes of Microfinance Over-Indebtedness: A Customer Protection Point of View," *Oxford Development Studies*, 41/sup1: S95–116.

Schicks, J., and R. Rosenberg (2011). "Too Much Microcredit? A Survey of the Evidence on Over-Indebtedness," *CGAP Occasional Paper*, 19.

Schilbach, F. (2019). "Alcohol and Self-control: A Field Experiment in India," *American Economic Review*, 109(4): 1290–1322.

Schuler, S. R., Lenzi, R., Badal S. H., and S. Nazneen (2018). "Men's Perspectives on Women's Empowerment and Intimate Partner Violence in Rural Bangladesh," *Culture, Health & Sexuality*, 20(1): 113–27.

Schurman, R. (2018). "Micro(soft) Managing a 'Green Revolution' for Africa: The New Donor Culture and International Agricultural Development," *World Development*, 112: 180–92.

Scott, J. C. (1977). *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. New-Haven: Yale University Press.

Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes for Improving the Human Condition Have Failed*. New-Haven: Yale University Press.

Scriven, M. (1991). *Evaluation Thesaurus* (4th ed.). Newbury Park, CA: Sage Publications.

Scriven, M. (2008). "A Summative Evaluation of RCT Methodology and an Alternative Approach to Causal Research," *Journal of Multidisciplinary Evaluation*, 5(9): 11–24.

Second International Study of Infarct Survival Collaborative Group (ISIS-2) (1988). "Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither among 17,187 Cases of Suspected Acute Myocardial Infarction," *Lancet*, 2: 349–60.

Sehon, S. R. and D. E. Stanley (2003). "A Philosophical Analysis of the Evidence-Based Medicine Debate," *BMC Health Services Research*, 3(1): 14.

Servet, J.-M. (2006). *Banquiers aux pieds nus*. Paris: Odile Jacob.

Servet, J.-M. (2011). "La crise du microcrédit en Andhra Pradesh (Inde)," *Revue Tiers Monde*, 3: 43–59.

Servet, J.-M. (2018). *L'Economie comportementale en question*. Paris: Fondation pour le Progrès de l'Homme.

Servet J.-M. and B. Tinel B. (2020). "The Behavioural and Neoliberal Foundations of Randomisations," *Strategic Change: Briefings in Entrepreneurial Finance*, 29(3): 293–299.

Shaffer, P. (2015). "Two Concepts of Causation: Implications for Poverty," *Development and Change*, 46(1): 148–66.

Shahar, E. (1997). "A Popperian Perspective of the Term 'Evidence-Based Medicine'," *Journal of Evaluation in Clinical Practice*, 3(2): 109–16.

Shand, D. G. (1975). "Propranolol," *New England Journal of Medicine*, 293: 280–4.

Shaw, L.W. and T. C. Chalmers (1970). "Ethics in Collaborative Clinical Trials," *Annals of the New York Academy of Sciences*, 169(2): 487–95.

Shelton, J. D. (2014) "Evidence-based Public Health: Not Only Whether It Works, But How It Can Be Made to Work Practicably at Scale," *Global Health: Science and Practice*, 2(3): 253–8.

Silverman, S. (2009). "From Randomized Controlled Trials to Observational Studies," *American Journal of Medicine* 122(2): 114–20.

Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation*. New York: Chapman; Hall.

Singh, I., Squire, L., and J. Strauss (eds.) (1986). *Agricultural Household Models: Extensions, Applications, and Policy*. Baltimore, MD: The Johns Hopkins University Press.

Skinner, Q. (2003). *Visions of Politics: Regarding Methods*. *Vol. 1 (2nd edition)*. Cambridge: Cambridge University Press.

Skoufias, E. and S. Parker (2001). "Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico," *Economía* 2(1): 45–86.

Snow, J. (1855). *On the Mode of Communication of Cholera*. London: Churchill. http://www.med.mcgill.ca/epidemiology/hanley/c609/material/SnowCholera/On_the_mode_of_communication_of_cholera.pdf

South African Cochrane Centre (2014). *Evidence-based Interventions for Diagnosing, Preventing and Treating Tuberculosis*. South African Cochrane Centre.

Spears, D., Ban, R., and O. Cumming (2020). "Trials and Tribulations: The Rise and Fall of the RCT in the WASH Sector." Chapter 6, this volume.

Spears, D., Ghosh, A., and O. Cumming (2013). "Open Defecation and Childhood Stunting in India: An Ecological Analysis of New Data from 112 Districts," *PloS one*, 8(9): e73784.

Squire, L. (1989). "Project Evaluation in Theory and Practice," in H. Chenery and T. N. Srinivasan, *Handbook of Development Economics*, Volume 2, Amsterdam: North-Holland, 1093–1137.

Squire, L. and H. van der Tak (1975). *Economic Analysis of Projects*, Baltimore and London: Johns Hopkins University Press for World Bank.

Staiger, D. and J. Stock (1997). "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65 (3): 557–86.

Stake, R. E. (2010). *Qualitative Research: How Things Work*. New York: Guilford Press.

Statistics Canada (2010). *Survey Methods and Practices*. Ottawa: Ministry of Industry.

Stenberg, K., Hanssen, O., Tan-Torres Edejer, T., Bertram, M., Brindley, C., Meshreky, A., Rosen, J.E., Stover, J., Verboom, P., and R. Sanders (2017). "Financing Transformative Health Systems towards Achievement of the Health Sustainable Development Goals: A Model for Projected Resource Needs in 67 Low-income and Middle-income Countries," *The Lancet Global Health*, 5(9): e875–e887.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., and B. Befani (2012). "Broadening the Range of Designs and Methods for Impact Evaluations." *Department for International Development Working Paper*, 38.

Stiglitz, J. (1986). "The New Development Economics," *World Development* 14 (2): 257–65.

Stiglitz, J. (2004). "The Post-Washington Consensus," *The Initiative for Policy Dialogue*.

Stiglitz, J. (2006). *Making Globalization Work*. New York: WW Norton.

Stiglitz, J. and A. Weiss (1981). "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 71: 393–410.

Stock, P. L. (1993). "The Function of Anecdote in Teacher Research," *English Education*, 25–3: 172–87.

Strassman D. and L. Polanyi (1995). "The Economist as Storyteller," in E. Kuiper and J. Sap (eds.), *Out of the Margin: Feminist Perspectives on Economics*. London: Routledge: 129–50.

Student (1938). "Comparison between Balanced and Random Arrangements of Field Plots," *Biometrika*, 29(3/4): 363–78.

Suppes, P. (1982). "Arguments for Randomizing," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982 (2): 464–75.

Svorenčik, A. (2015). *The Experimental Turn in Economics: A History of Experimental Economics*. Utrecht School of Economics Dissertation Series, 29.

Tarozzi, A., Desai, J., and K. Johnson (2015). "The Impacts of Microcredit: Evidence from Ethiopia," *American Economic Journal: Applied Economics*, 7(1): 54–89.

Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L., and J. Yoong (2014). "Micro-loans, Insecticide-treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India," *American Economic Review*, 104(7): 1909–41.

Tarp, F. (2009). *Aid Effectiveness*, United Nations University, WIDER, Helsinki.

Tavernise, S. (2015). "Few Health System Studies Use Top Method, Report Says," *New York Times*, February 12.

Taylor, K. M, Margolese, R. G., and C. L. Solskolne (1984). "Physicians' Reasons for Not Entering Eligible Patients in a Randomized Trial of Surgery for Breast Cancer," *New England Journal of Medicine*, 310: 1363–7.

Taylor, M. (2011). "'Freedom from Poverty Is Not for Free': Rural Development and the Microfinance Crisis in Andhra Pradesh, India," *Journal of Agrarian Change*, 11/4: 484–504.

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, S., and P. Garner (2015). "Deworming Drugs for Soil-transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance," *The Cochrane Database of Systematic Reviews*, 2015(7): CD000371–CD000371.

Teele, D. L. (ed.) (2014). *Field Experiments and Their Critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven & London: Yale University Press.

Tendler, J. (1993). *New Lessons from Old Projects: The Workings of Rural Development in Northeast Brazil. A World Bank Operations Evaluation Study*, Washington, D.C.: World Bank,

Thaler, R. H. (2015). *Misbehaving. The Making of Behavioral Economics*, New York: Penguin.

Theo, T. (2009). "Philosophical Concerns in Critical Psychology," in D. Fox, I. Prillentensky, and S. Austin (eds.), *Critical Psychology: An Introduction*, Sage Publications, Thousand Oaks, 44.

Theroux P., Franklin D., Ross, J. Jr., and W. S. Kemper (1974). "Regional Myocardial Function during Acute Coronary Occlusion and Its Modification by Pharmacologic Agents in the Dog," *Circulation Research*, 35: 896–908.

Thomson, D. R., Amoroso, C., Atwood, A., Bonds, M. H., Cyamatare Rwabukwisi, F., Drobac, P., Finnegan, K. E., Farmer, D. B., Farmer, D. B., Farmer, P. E., and A. Habinshuti (2018). "Impact of a Health System Strengthening Intervention on Maternal and Child Health Outputs and Outcomes in Rural Rwanda 2005–2010," *BMJ Global Health*, 3: e000674.

Tinbergen, J. (1956). *Economic Policy: Principles and Design*, Amsterdam: North Holland.

Todd, P. and K. Wolpin (2006). "Assessing the Impact of a School Subsidy Program in Mexico using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling," *American Economic Review*, 96(5): 1384–1417.

Ubel, P. A. and R. Silbergleit (2011). "Behavioral Equipoise: A Way to Resolve Ethical Stalemates in Clinical Research," *American Journal of Bioethics*, 11(2): 1–8.

United Nations (2015). *The Millennium Development Goals Report*, New-York: United Nations.

United Nations (2020). *Private Debt and Human Rights* (Report of the Independent Expert on the effects of foreign debt and other related international financial obligations of States on the full enjoyment of human rights, particularly economic, social and cultural rights). Geneva: United Nations.

UNDP (1999). *Human Development Report*. New York, Oxford: Oxford University Press

van der Meulen Rodgers, Y., Bebbington, A., Boone, C., Dell'Angelo, J., Platteau, J.-P., and A. Agrawal (2020). "Experimental Approaches in Development and Poverty Alleviation," *World Development*, 127 104807.

Vass, M. (2010). "Prevention of Functional Decline in Older People: The Danish Randomised Intervention Trial on Preventative Home Visits." Doctoral Dissertation, Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark.

Veatch, R. M. (2007). "The Irrelevance of Equipoise," *Journal of Medicine and Philosophy*, 32(2): 167–83.

Vedung, E. (2010). "Four Waves of Evaluation Diffusion," *Evaluation*, 16: 263–77.

Victora, C. G., Black, R. E., Ties Boerma, J., and J. Bryce (2011) "Measuring Impact in the Millennium Development Goal Era and Beyond: A New Approach to Large-scale Effectiveness Evaluations," *The Lancet*, 377(9759): 85–95.

Vivalt, E. (2019). "Specification Searching and Significance Inflation across Time, Methods and Disciplines," *Oxford Bulletin of Economics and Statistics*, 81(4): 797–816.

Vivalt, E. (2020). "Using Priors in Experimental Designs: How Much Are We Leaving on the Table?" Chapter 11, this volume.

Vivalt, E. (forthcoming). "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association*. https://doi.org/10.1093/jeeasn/jvaa019

Vivalt, E. and A. Coville (2016). "How Do Policymakers Update?" Unpublished manuscript.

Vrieze, J. de (2018). "The Metawars," *Science*, 361(6408): 1184–8.

de Waal, A. (1997). *Famine Crimes: Politics and the Disaster Relief Industry in Africa*. Melton: James Currey.

Webber, S. and C. Prouse (2018). "The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development,'" *Economic Geography*, 94(2): 166–87.

Weber, M. (1958). "Science and Vocation," in H. H. Gerth and C. Wright Mills (eds.), *Max Weber: Essays in Sociology,* New-York: New York University Press, 155.

Weijer, C., Glass, K. C., and S. H. Shapiro (2000). "Why Clinical Equipoise, and Not the Uncertainty Principle, Is the Moral Underpinning of the RCT," *British Medical Journal*, 321: 756–8.

Whidden, C., Kayentao, K., Liu, J. X., Lee, S., Keita, Y., Diakité, D., Keita, A., Diarra, S., Edwards, J., Yembrick, A., Holeman, I., Samaké, S., Plea, B., Coumaré, M., and A. D. Johnson (2018). "Improving Community Health Worker Performance by Using a Personalised Feedback Dashboard for Supervision: A Randomised Controlled Trial," *Journal of Global Health*, 8(2): 020418.

White, H. (2013). "An Introduction to the Use of Randomised Control Trials to Evaluate Development Interventions," *Journal of Development Effectiveness*, 5(1): 30–49.

White, H. (2014). "Ten Things that Can Go Wrong with Randomised Control Trials," Evidence Matters Blog, International Initiative for Impact Evaluation. https://www.3ieimpact.org/blogs/ten-things-can-go-wrong-randomised-controlled-trials. Last accessed November 12, 2019.

White, H. and E. Masset. (2018). "The Rise of Impact Evaluations and Challenges which CEDIL Is to Address," *Journal of Development Effectiveness*, 10 (4): 393–9. doi.org/10.1080/19439342.2018.1539387.

Whittle, D. (2011). "If Not Randomized Trials, Then What?," *Pulling for the Underdog Blog*, June 1, 2011. https://www.denniswhittle.com/2011/06/randomized-trials-not-silver-bullet.html. Last accessed November 12, 2019.

Wilke, A. and M. Humphreys (2019). "Field Experiments, Theory and External Validity," Working Paper, https://www.dropbox.com/s/47s52xv0frrnvml/20190703_Wilke_Humphreys.pdf?dl=0

WMA General Assembly. (2014). *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*, 9th revision (1st version: 1964). Fortaleza: World Medical Association.

Woolcock, M. (2013). "Using Case Studies to Explore the External Validity of 'Complex' Development Interventions," *Evaluation*, 19(3): 229–48.

World Bank (2005). *World Development Report—Economic Growth in the 1990s: Learning from a Decade of Reform*. Washington.

World Bank (2012). *World Bank Group Impact Evaluations: Relevance and Effectiveness,* Washington DC: Independent Evaluation Group, World Bank.

World Bank (2015). *World Development Report: Mind, Society, and Behavior,* Washington DC: IBRD/WB.

World Bank (2016). *Transforming Development through Impact Evaluation, I2i DIME Annual Report*, Washington DC: World Bank.

World Health Organization (2010). *Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and their Measurement Strategies*. Geneva: WHO.

World Health Organization and the World Bank (2017). *Tracking Universal Health Coverage: 2017 Global Monitoring Report*. Geneva: WHO/IBRD/WB.

Worrall, John (2007). "Why There's No Cause to Randomize," *The British Journal for the Philosophy of Science*, 58(3): 451–88.

Wrong, Michela (2009). *It's Our Turn to Eat: The Story of a Kenyan Whistleblower*. New York: Harper.

Wu, D. (1973). "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, 41(4): 733–50.

Wydick, B. (2016). "Microfinance on the Margin: Why Recent Impact Studies May Understate Average Treatment Effects," *Journal of Development Effectiveness* 8(2): 257–65.

Wydick, B. (2018). "Review of Randomistas: How Radical Researchers Changed Our World," *Development Impact Blog, World Bank*. May 21, 2018. https://blogs.worldbank.org/impactevaluations/review-randomistas-how-radical-researchers-changed-our-world. Last accessed November 12, 2019.

Yang, Y. (2019). "The Open Secret of Development Economics," Project Syndicate, October 22.

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., Gerland, P., New, J. R., and L. Alkema (2015). "Global, Regional, and National Levels and Trends in Under-5 Mortality between 1990 and 2015, with Scenario-based Projections to 2030: A Systematic Analysis by the UN Inter-agency Group for Child Mortality Estimation," *The Lancet*, 386(10010): 2275–86.

Young, A. (2019). "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics*, 134(2): 557–98.

Young, J., Harrison, J. White, G., May, J., and M. Solomon (2004). "Developing Measures of Surgeons' Equipoise to Assess the Feasibility of Randomized Controlled Trials in Vascular Surgery," *Surgery*, 136(5): 1070–6.

Yusuf, S., Peto, R., Lewis, J., Collins, R., and P. Sleight (1985). "Beta Blockade During and After Myocardial Infarction: An Overview of the Randomized Trials." *Progress in Cardiovascular Diseases*, 27(5): 335–71.

Ziliak, S. T. (2014). "Balanced versus Randomized Field Experiments in Economics: Why W. S. Gosset aka 'Student' Matters," *Review of Behavioral Economics*, 1: 167–208.

Ziliak, S. T. and E. R. Teather-Posadas (2016). "The Unprincipled Randomization Principle in Economics and Medicine," in G. DeMartino and D. McCloskey (eds.), *Oxford Handbook on Professional Economic Ethics*, New-York and Oxford: Oxford University Press. Online version https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199766635.001.0001/oxfordhb-9780199766635-e-44