© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

INTRODUCTION

WHY FAILURES?

Thomas Edison tried hundreds of materials before discovering that bamboo fiber could serve as a lightbulb filament. When asked how it felt to fail so much, he reportedly answered, "I have not failed 700 times. I have not failed once. I have succeeded in proving that those 700 ways will not work. When I have eliminated the ways that will not work, I will find the way that will work."

We salute Edison and, in creating this book, we drew inspiration from his attitude. Like the lightbulb, knowledge is a product; research produces it. Just as Edison tinkered with his filaments to see which produced a bright and lasting light, researchers tinker with research designs to see which produce meaningful and reliable knowledge. The more experimenters share their private failures so that others can learn, the more quickly we will find the ways that work. This book is an effort to document some of the many research failures in this space so that others, to paraphrase Edison, do not have to re-prove that all those 700 ways do *not* work.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

In collecting and reflecting on failures we have vacillated between two guiding thoughts. The first, like the Edison story, is inspirational:

Failure is the key to success; each mistake teaches us something.
—MORIHEI UESHIBA

Some clichés actually contain good advice. This one happens to be our core motivation for writing this book.

The second mantra, while less sanguine, is no less true:

Failure: When Your Best Just Isn't Good Enough
—WWW.DESPAIR.COM/PRODUCTS/FAILURE

Every basic textbook in economics teaches why people should ignore sunk costs when making decisions. Yet reality of course is a tad different: people often let sunk costs sway decisions. We must remember that we will all fail here or there, despite our best efforts. Sometimes we just cannot do better. Rather than double down, we may need to learn and move on.

SUCCESS; OR WHY THE FAILURES ARE WORTH THE HASSLE

Today, conversations about poverty alleviation and development are much more focused on evidence than they were before—a shift due, in large part, to the radical drop in the price of data and the growth of randomized controlled trials (RCTs) in the development sector. Long the dominant methodology for determining medical treatments, RCTs found their way into domestic social policy discussions as early as the 1960s when they were used to evaluate government assistance programs, such as negative income tax rates for the poor. In the 1990s, a new crop of development economists began using RCTs in the field to evaluate aid programs. Their work generated enthusi-

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

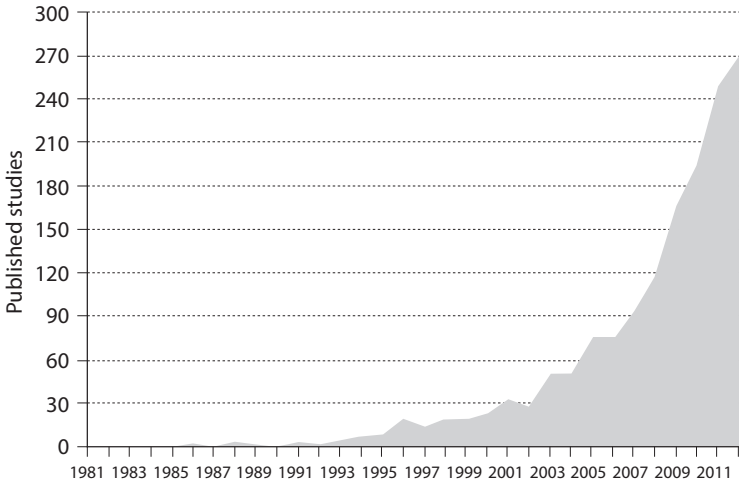

FIGURE 1. Published randomized controlled trials in development, 1981–2013. Drew B. Cameron, Anjini Mishra, and Annette N. Brown, "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* 8, no. 1 (2016): 1–21.

asm for the RCT approach and helped spark a pervasive trend, illustrated in figure 1 with data compiled by the International Initiative for Impact Evaluation (3ie), a donor consortium.

PUBLISHED RANDOMIZED CONTROLLED TRIALS IN DEVELOPMENT, 1981—2013

The role of RCTs in development has expanded not just in volume but also in variety. Many early applications were straightforward program evaluations ("What impact did program X have on outcomes Y and Z?"). Now RCTs are commonly used to test specific theories in development, such as whether cognitive attention influences savings and borrowing behavior; whether specific forms of moral hazard are present in credit markets; how social networks influence the adoption of new agricultural

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

technologies among smallholder farmers; whether increasing attention or information can lead to people taking their medication more reliably; and whether a commitment device to invest in fertilizer improves yields for farmers. Finally, RCTs help address operational issues, such as the optimal repayment schedule for a loan, how to price sanitation services, or how to describe cash transfer programs to recipients. Naturally these are not mutually exclusive categories: some RCTs examine the efficacy of a program while successfully testing a theory and answering important operational questions. Such is the aim of many studies, but there is variation in how well a study accomplishes each of the three.

As the number and range of RCTs have grown, so have the quantity and quality of lessons learned. Consequently, policies are changing, too—often slowly and with a lot of pushing, but changing nonetheless. Innovations for Poverty Action (IPA, a nonprofit organization that Dean founded) and MIT's Abdul Latif Jameel Poverty Action Lab (J-PAL) collaborate and keep a tally of such victories from their own collective work, called "scale-ups": cases where social programs were expanded after rigorous evaluation produced strong evidence of effectiveness. The work spans many areas, including education, health, microfinance, income support, and political economy. Here are eight examples:

- Improved targeting and distribution of subsidized rice—65+ million people[1]
- Remedial education with volunteer tutors—47+ million children[2]
- Conditional community block grants—6+ million people[3]
- School-based deworming—95+ million children[4]
- Chlorine dispensers for safe water—7+ million people[5]

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

- Free distribution of insecticidal bed nets for malaria (*no figure available*)[6]
- Police skills-training to improve investigation quality and victim satisfaction—10 percent of police personnel in Rajasthan, India[7]
- Integrated "Graduation" grants for training, coaching, and financial service programs that aim to increase the income of those in extreme poverty—400,000+ households[8]

The diversity of scale-ups is telling.

On one end of the scale-ups spectrum are tweaks to large, established programs. In these cases, studies found subtle changes to the design or delivery that drastically improved overall program effectiveness. For example, researchers worked with the Indonesian government to add a simple performance incentive to a community block grant program such that communities that achieved better health and education outcomes would receive larger grants for the following year. This drove significant health improvements compared to similar grant amounts given without incentives.[9]

A study of a Colombian government-run conditional monthly cash transfer program found big impacts from *delaying* a portion of the monthly payment until school fees were due the following term. This subtle variation produced the same short-term gains in attendance as did immediate payments but also significantly increased the portion of children who continued on to the next grade. Similarly, delaying payment until high school graduation maintained short-term attendance gains while increasing matriculation in postsecondary education.[10]

Again in Indonesia, another program took on systematic underuse of the government's largest social program, a rice subsidy for its poorest citizens. Only 30 percent of eligible

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

households knew about the program, and among those who did, many were confused about the price and the subsidy amount to which they were entitled. An evaluation tested the impact of two interventions: distributing personal ID cards that included price and quantity information and served to remind people they were eligible for the subsidy; and a community-based awareness campaign. Both proved highly effective in getting eligible people to use the program and in combating overcharging. For an additional cost of about $2.40 per household, they increased the average value of the subsidy by more than $12.[11]

Finally, insecticidal bed nets have long been used to fight malaria, but until recently there was no consensus about pricing. Some felt they should be given away for free—this way nobody would be priced out of protection. Others felt recipients should pay *something*, reasoning that a free handout would more likely be forgotten or neglected than would an intentional investment. Evaluation provided critical guidance: When offered for free in Kenya, bed nets reached far more people and were equally likely to be used properly.[12] Free distribution is now endorsed by the UK's Department for International Development, the UN Millennium Project, and other leading organizations.

On the other end of the scale-ups spectrum are big program evaluations that then lead to entire programs being scaled. Three examples are remedial education with volunteer tutors,[13] chlorine dispensers for safe water,[14] and an integrated multifaceted grant program (often called a "Graduation" program) to increase income for those in extreme poverty.[15] In these cases, the initial evaluations were focused on trying something new or something without solid evidence rather than fine-tuning existing approaches. These programs began as limited-scale pilots and grew as repeated rounds of evaluation confirmed their effectiveness.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Remedial education with tutors began as a 200-school pilot in the Indian state of Gujarat and now reaches nearly 34 million students across the country. Taken together, the scale-ups on IPA and J-PAL's list reach over 200 million people in the developing world.

Chlorine dispensers for safe water address a simple and pervasive problem. Adding a few drops of chlorine is a cheap, effective, and widely available way to sterilize drinking water that might otherwise make people sick; but usage rates are low, even among those who know about chlorine disinfectant, have access, and would benefit from using it. A potential solution sprang from the behavioral insight that people might be more likely to use chlorine if it were available "just-in-time," at the moment when they collected the water. Voilà: Chlorine dispensers were installed next to wells and community water taps. This also lowered cost, as there was only one end-node per village that needed replenishing with chlorine. A series of evaluations confirmed the positive health impact of the dispensers and helped drive scale-up. As of the end of 2015, some 7.6 million individuals in East Africa use clean water thanks to this simple solution.

A more recent example is cash transfer programs that give money directly to the poor to use however they see fit. Initially such programs were met with skepticism. Many wondered whether the poor would use the money wisely. But an RCT of a cash transfer program in Kenya, administered by the nonprofit GiveDirectly, found large impacts on recipients' incomes, assets, food consumption, psychological well-being, and more.[16] Encouraged by the evidence, the philanthropic funding organization Good Ventures gave $25 million to GiveDirectly to expand its programs. Based on these early results, other countries are expanding their own cash transfer programs. Many questions remain about how best to design these programs, but they do provide a useful benchmark for comparison. We

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

might now ask of any poverty alleviation program that requires funding, "Does it do better than just giving that money directly to the poor?"

Cash transfers could be thought of as what-to-do-when-you-do-not-know-what-else-to-do. But they are not a panacea; many problems of poverty are not simply a by-product of having less money. Had this been true, redistributive policies would have solved many problems long ago. There are likely many market failures present: low information for the poor or lack of access to markets, for example. This implies that efforts that tackle market failures, perhaps along with redistribution, may have a multiplicative effect that mere redistribution does not have. In other words, transfer $1 of cash and the recipient gets $1 (actually about $0.95, which is quite good). But maybe there are ways of transferring $1 along with $0.50 of services that then generates $2 for the recipient. That is a bigger bang for your buck (or buck-fifty). To examine that, we turn to the last example, a "Graduation" program designed to help build income for families in extreme poverty.

Researchers from IPA and J-PAL completed a six-country evaluation of a Graduation program that provides families with comprehensive assistance, including a productive asset (e.g., four goats), training to care for the asset, basic nutrition and health care support, access to savings, and life-skills coaching. The theory is that such coordinated, multilayered support helps families graduate to self-sufficiency. The evaluation followed some 21,000 of the world's poorest people, over three years, and found strong positive social returns ranging from 133 to 433 percent. That is, the program generated between $1.33 and $4.33 in increased consumption for the household for each dollar spent.

The Graduation study is a breakthrough in scale and coordination. Following individuals over three years is a challenge

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

in itself. Doing so across six countries (Ethiopia, Ghana, Honduras, India, Pakistan, and Peru); asking comparable questions in six languages; measuring and comparing social and financial outcomes across six economies with regional currencies, prices, and staple goods; and working with six different implementation teams was an even bigger undertaking.

Replications usually happen one by one, if at all. One successful study might generate interest (and funding) for a replication, so it often takes years to build up a body of evidence around a particular program or approach. With the Graduation study researchers have evidence from six settings all at once (and a seventh, from Bangladesh, conducted by separate researchers), enabling them to see differences across contexts. It brings us closer to understanding *how* Graduation programs truly work—and for whom and under what conditions they might work best.

This all makes the Graduation study sound like a blistering success, but there were failures here, too. There were some simple "oops" mistakes: for example, leaving out a module in a survey, and then having no measure of mental health in a round of data. Many outcome measures were, upon reflection, not as comparable across sites as researchers would have liked. Other challenges were inevitable consequences of differences across countries, say, in researchers' abilities to ask detailed questions about the costs of rearing an animal. Finally, each site made huge promises to deliver detailed monitoring data in order to help capture contextual differences between the six settings, but none of these data became available in the end, rendering comparisons across sites more challenging.

Academic journals and media reports are now full of success stories of rigorous evidence being brought to the fight against poverty. IPA and J-PAL have been featured in major news outlets of all stripes, from *Time*, TED, and the *New Yorker*

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

to the *Economist* and the *Wall Street Journal*. Despite all the good press, not everything is rosy. We believe sharing the studies that are not fit for the *New York Times*—the ones that will otherwise just die, undocumented, in the memory of the researchers—can yield valuable lessons.

We also had a deeper motivation in writing this book: Researchers must begin talking about failures to ensure evidence plays an *appropriate* role in policy. The language of RCTs as the "gold standard" for evidence has no doubt helped fuel their rise. But it is telling that one rarely hears RCT researchers make such claims; there is a risk of overreaching. The fact is, where it is appropriate to use, a well-executed RCT will provide the strongest evidence of causality. It will say, more decisively and precisely than other methods, whether program X caused outcome Y for the population being studied. Much of this nuance gets lost in public discussion. All too often, what comes through is just that RCT equals gold standard. But the "where it is appropriate" and "well-executed" parts are equally important! Not every program or theory is amenable to study by an RCT; even when one is, the RCT can be poorly executed, producing no valuable knowledge. (Read on for many such examples.)

In fact, there is meta-evidence that employing poor methods is correlated with biased results. In an analysis of medical RCTs, researchers found that studies with either inadequate or unclear concealment of assignment to treatment and control had larger treatment effects. This is a striking result that has one obvious and one less-than-obvious interpretation. The obvious: poor methods lead to higher variance in outcomes or, worse, suggest possible manipulation. Thus they generate a higher proportion with statistically significant treatment effects, and in general larger treatment effects get published. The less-than-obvious: perhaps journal editors and referees are more likely to overlook sloppy methods when the treatment effects are large.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

They are inclined to be fastidious about borderline results but do not want nitpicky details to get in the way of reporting "big" results. Alas, to test between these two explanations one ideally would randomly assign the use of sloppy methods to journal submissions to see whether the sloppy methods affected acceptance rates differentially for high and low treatment effect papers—hardly a feasible study.

Bottom line, a bad RCT can be worse than doing no study at all: it teaches us little, uses up resources that could be spent on providing more services (even if of uncertain value), likely sours people on the notion of RCTs and research in general, and if believed may even steer us in the wrong direction. Naturally the same can be said about any poorly designed or poorly executed research study.

OUR FOCUS IN THIS BOOK

Even in the relatively confined space of international development research, there are many different kinds of failures. We will not attempt to address, or even describe, them all. Let us say more about what we will and will not discuss.

What we do focus on in this book are *research failures*, cases where a study is conceived with a particular question in mind (e.g., "Does financial literacy training help borrowers repay their microloans?") but does not manage to answer it. Why might this happen? Such cases fall into two categories. Either researchers started out with a faulty plan, or they had a good plan but were derailed by events that took place once the study was underway. The kinds of failures we discuss in chapters 1 and 2, research setting and technical design, fall into the first category. The second category includes partner organization challenges, survey and measurement execution problems, and low participation rates, which we discuss in chapters 3, 4, and 5.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

In contrast, sometimes a product or service (e.g., microloans, scholarships, vaccinations) is delivered as planned, and then a well-executed evaluation returns a precise null result—statistical confirmation that the intervention did not causally influence target outcomes. Simply put, the program did not work. We call such cases *idea failures.* These are important lessons to learn and to share. But they are not our subject here.

Many idea failures are actually research successes. With a precisely estimated result of "no impact" in hand, one can confidently move on and try other approaches. The only real failure we see in these scenarios is with the academic publishing system: precise "no impact" results are often rejected by scholarly journals, an embarrassing and pernicious reality. Many are trying to address this, but, to mix bad metaphors, it is a bit like trying to herd cats in an effort to move Sisyphus's rock. A good aspiration for an evaluation is to shed light on *why* something works, not merely whether it works. Idea failures can be just as revealing as successes: if the approach seemed sensible, why did it not work? "No impact" results notwithstanding, if a study can explain that, then the publishing system should reward it just as it rewards research documenting ideas that work.

WHAT IS IN THE REST OF THE BOOK

In part 1 (chapters 1–5) we sort research failures into five broad categories, describing how and why they might arise and highlighting common instances or subtypes of each. The categories are as follows: inappropriate research setting, technical design flaws, partner organization challenges, survey and measurement execution problems, and low participation rates. Occasionally we illustrate these with hypotheticals (e.g., "Imagine you are evaluating a school feeding program in 200

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

schools . . ."), but where possible we include actual examples from the field.

In part 2 (chapters 6–11) we present case studies of six failed projects in more detail, including background and motivation, study design and implementation plan, what exactly went wrong, and lessons learned. These are forensic-style reports on projects that genuinely flopped. Like most instances that call for forensics, the cases they describe are messy and multifaceted, with each touching on multiple failure types. Indeed, one lesson that emerged as we wrote and compiled these cases is that individual failures tend to snowball quickly, especially if they are not caught and addressed immediately. It is worth mentioning that many of the cases in part 2 deal with "microcredit plus" programs—that is, programs that seek to bundle some additional services along with microloans. This similarity across cases is partly a consequence of the simple fact that we (particularly Dean) have done a lot of research on microfinance. But while these cases are programmatically similar, each one failed in its own special way and for its own special reasons. This is, in itself, an interesting finding: the potential pitfalls of research depend not just on the program being studied but on many other factors, too. Conversely, studies of vastly different kinds of programs can fail in similar ways and for similar reasons.

Finally, the conclusion and appendix distill key lessons and themes from across the cases discussed in parts 1 and 2 and offer some guidance for those about to embark on their own field studies. By this point, you will have read a fair bit about what not to do. But what *do* you do? The advice there hardly amounts to a comprehensive guide, but it is a start. In addition to providing positive takeaways from cases in this book, we also point to external resources that provide concrete guidelines,

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

procedures, and strategies to help researchers design and run successful field studies.

WHAT IS BEYOND THIS BOOK

We are glad and grateful that dozens of colleagues, from first-time research assistants to tenured professors, have joined in this effort, publicly sharing their own difficult experiences so that others can benefit. But alas, the field of international development research lacks a tradition of talking openly about failures. Finding willing contributors was difficult—which is one reason why many of the examples we discuss here are our own.

In writing this book, we reached out to more researchers than are featured in these pages. Many admired the venture but had no story they were willing to share publicly. Some had concerns about the sensitivities of funders or partner organizations that had been involved in failed research efforts (especially if they hoped to work with those funders or partners again in the future); others were reluctant to cast as "failures" studies that had significant hiccups but still produced some publishable results. Perhaps some simply did not want their name on a list of researchers who made mistakes.

Given the realities of funding, publication, and career advancement these researchers face, their reasons are understandable. It is the expectations placed on them that are unrealistic. Anyone who tries enough times will fail occasionally; why pretend otherwise? The subject need not be taboo. If more researchers make a habit of sharing their failures, the quality of research can improve overall.

This book represents a first step down that path. Beyond the examples in these pages, we are launching, with David McKenzie and Berk Özler of the World Bank, an online companion—an effort to aggregate and categorize failure stories

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

on their moderated blog. Ideally it will become common practice among researchers, when studies go bad, to briefly capture what went wrong and share it with the community. If this takes off, we will then work with them to launch the failure repository as a stand-alone site. Equally important, we hope researchers embarking on new studies genuinely use the resource to look at past failures, incorporate relevant lessons, and avoid learning the hard way themselves.

For general queries, contact webmaster@press.princeton.edu