

EXPERIMENTAL CONVERSATIONS

**Perspectives on Randomized Trials
in Development Economics**

Timothy N. Ogden,
editor

9 Lant Pritchett

Lant is professor of the Practice of International Development at the Kennedy School of Government at Harvard, and a senior fellow at the Center for Global Development and Bureau for Research and Economic Analysis of Development. Lant spent more than a decade in various research roles at the World Bank. He is author of *Let Their People Come* and *The Rebirth of Education: Schooling Ain't Learning*.

TO: As I read what you've written on RCTs, I see two streams of critique. One is, "we're focusing on things that are too small." The second is the overhyping of certainty as a result of the methodology.

LP: I would say there are three critiques.

One critique is you're tackling only small problems because you're tackling problems almost by construction that are individuated. Meaning, the level I'm intervening at is the person,¹ whereas what we economists know and specialize in is how systems work.

The only insight that economists bring to the world is that it's not at the individual level that success and failure happens. The interaction between individuals is what matters. That's the insight of Adam Smith right? The market as a system mediates the actions of self-interested individuals to produce teleologically unintended but normatively rankable outcomes. Meaning, to rephrase Adam Smith, the butcher doesn't give you meat because he likes you, he does it for money but that produces actually really good outcomes that nobody intended.

1. See interview with Frank DeGiovanni (chapter 13), where he comments on only funding RCTs for programs that operate on an individual level because community-level interventions are too hard or too expensive to evaluate with RCTs.

So in RCTs you've reduced the system to the individual. Nothing super-important about development happens at the individual level.²

The second critique is the RCT folks have the fundamental political economy of policy change wrong.^{3[1]} Think of any development activity as being a mapping from inputs to activities to outputs to outcomes. The claim the randomistas make is we're going to take cash, we're going to turn it into the activity of doing RCTs, and that's going to lead to better outcomes. But they just don't have a coherent, much less evidence-based, causal model of how the outputs they produce are going to yield the outcomes they intend.⁴ That requires some positive theory of policy: how do policies actually get made, adopted, and implemented? They have this just unbelievably Cro-Magnon simple model of policy adoption that essentially asserts that once there is better knowledge in the world about policies, that will lead to better policies being adopted and implemented.

You could make an argument like that, but it's just wrong. So they're wrong about the fact that the outputs of RCTs are going to change policy.⁵ What's amazing is that precisely their critique of all other development endeavors is that they were wrong about their causal model of outputs to outcomes. Their attack on microcredit is that while it's true you can turn inputs into loans and those loans produce activities that are outputs, you were wrong about the causal model by which the availability of microcredit would transform outcomes for individuals.

2. For more on Pritchett's thinking on "small" interventions in the history of development, see his post on the Center for Global Development blog, "Is Your Impact Evaluation Asking Questions That Matter?," <http://www.cgdev.org/blog/your-impact-evaluation-asking-questions-matter-four-part-smell-test>, where he argues that few, if any, of the interventions studied most via RCTs played any significant role in the development of wealthy countries.

3. For additional background to Pritchett's point here, see "The Policy Irrelevance of the Economics of Education," a paper prepared as part of a Brookings conference titled, "What Works in Development? Thinking Big and Thinking Small." The collected papers from the conference have been published in a volume with the same name.

4. Most of the interviews include a question about the individual's theory of change and their experience in their work driving policy or product changes.

5. See interview with Frank DeGiovanni (chapter 13), where he argues that RCTs have been necessary to influence policy makers to consider "graduation model" packages of interventions for ultra-poor households.

It's their most powerful critique of what everyone else is doing, and yet they don't have a coherent answer about what they are doing. What is their coherent model that moves from the outputs of the research papers to the outcome of better lives for poor people? That has to be a political economy model. A political model if it's public policy you think is changing, or an organizational model if you think you're going to change the behaviors of organizations. It's just so frustrating because you can't even have this debate with them. They don't even pretend to have a coherent model of this. And yet it's precisely the criticism they levy against everyone else.

The third critique is they are just deeply wrong about the science and what science means and how science works. Meaning they're wrong about the relative importance of internal versus external validity.⁶ There are multiple levels to that critique. But they're just obsessed with internal validity and their basic response to the external validity problem, since they started, was, "well we'll deal with that later, since we have no internally valid estimates of program effects. Why are you worried about whether the ones we have are externally valid?" A bunch of us were saying there's not going to be any external validity. This is not going to add up to anything because there's huge heterogeneity in program impacts based on nonrigorous methodological studies.^{7[2]} Now the future came and they've just replicated exactly the same variance we already knew was there. If you're a policy maker and want to know if I undertake a program of this type, what will its impact be in my context, they have contributed exactly zero to answering that question. People are no better off in terms of knowledge than they were before.

Everyone who was a practitioner of any type knew external validity was a huge problem, and part of the reason the RCT skeptics pooh-poohed the internal validity obsession was because they knew it wasn't that big of a deal. Now with the randomistas own data we can show that we were right and they were wrong.

6. See interview with Jonathan Morduch (chapter 4) for more discussion of the history and importance of internally valid estimates of impact.

7. An example can be found in Pritchett's paper "What Education Production Functions Really Show." While the paper is focused on a model explaining the allocation of education spending and ways to improve the effectiveness of education spending, it illustrates how broad the range of program impact estimates was in studies of marginal products of education investment.

We now have six estimates of the impact of microcredit. If we were going to predict microcredit's impact in some context, would we be better off with the nonexperimental internally invalid estimate from our own context or an RCT from another context? The answer is you would increase your prediction error by using the RCT over your own context's bad estimate.^{8[3]}

I don't see how the debate can be won any more conclusively than that. What they are proposing as a scientific approach leads to significantly worse prediction errors than the stupidest possible alternative of just relying on the existing OLS⁹ estimates in context.

TO: Isn't there an aspect of the external validity question that comes down to whether we are learning things about contexts or about human beings and behavior? Haven't we learned something about human behavior?¹⁰ Something similar to the systematic biases from Kahneman's research?^{11[4]}

LP: I disagree there. An economist might say that all human beings seek to maximize utility right? That's true but it's also contentless. Because what maximizes utility in two different contexts might be completely different. Proving that human beings maximize utility in lots of different contexts doesn't answer any questions about the absolute magnitude of program impacts unless you also argue that what is in fact in the utility function is the same and roughly to the same degree in all contexts. There's no evidence for that. Economists have never asserted that.

8. In two papers co-authored with Justin Sandefur of the Center for Global Development, Pritchett argues that the differences in effect size from one context to another dwarf the differences between randomized evaluations and nonrandomized evaluations of impact in the same context. Therefore a more accurate estimate of effectiveness can be found by relying on a nonrandomized evaluation in the same context than a randomized evaluation from a different context. Specifically here, he refers to an analysis of the papers in the January 2015 issue of the *American Economics Journal: Applied Economics*.

9. Ordinary least squares is an econometric approach to assessing how one variable influences another where treatments have not been randomized.

10. A version of this same question is posed to many of the interviewees, most of whom believe that RCTs have been particularly effective at uncovering common human behaviors and that these insights are particularly useful.

11. Daniel Kahneman, who shared the 2002 Nobel Prize in economics despite being a social psychologist, pioneered the exploration of cognitive biases and their effect on decision making, with important implications for economics.

Our whole point is that people have different utility functions. That's why trade exists. The point of EC101, the first model you learn, is that people trade because they have different preferences and that leads to higher welfare for both people.

When anthropologists run the same experiment in different contexts, they get very different results. So as far as I know, one of the few things we know scientifically across lots of cultures is playing the ultimatum game. And if you play the ultimatum game across a lot of different cultures, you get extraordinarily different results.¹² The idea that there is external validity in program impacts but not in something simple like the ultimatum game just seems loopy to me.

Speaking of having it both ways, the argument that there will be external validity because there are universal features of human beings kind of runs up against the issue of working on "big" versus "small problems." If these heuristic biases are universal or nearly so, then it is hard to make them the explanation of huge variations across countries in development outcomes—either across countries or over time. If these Kahneman-like systematic flaws in rationality are a barrier to development, then why are the USA and Japan and Europe developed?

Now let's go back to basic science. Physics 101. If we write down a Newtonian law of motion, we do it because we know that there is invariance. If we translate an experiment to some other space, we're going to get exactly the same result. That's a classic example of an invariance law. There are certain aspects of medicine where we have invariance. We kind of know how cells behave. An invariance law can also specify ways in which the outcome will vary—because of temperature, for example. So we can do experiments on cells from all over the world and expect the results to be the same.

But that's precisely what we don't have in economics. What we do have says that people and systems are not invariant. So, for instance, no one ever expected that the impact on a person of having the offer of microcredit at a given interest rate would be constant in some way. That's just absurd. The fundamentals of how to do science somehow got lost in the enthusiasm for RCTs.

TO: Why has the RCT movement captured so much attention and energy?

12. See interview with Dean Yang (chapter 12) where the use of the ultimatum game, sometimes referred to as the dictator game, is discussed.

LP: I think there were three separate trends. The success of any social movement usually takes a confluence of things. It's often the case that many of the elements existed for a long time. There had been a civil rights movement since the end of the Civil War at least, so why did it suddenly have success in the 1960s and not before?

First, the RCT movement was in some sense the logical extension of debates about identification¹³ that had been going on since the Cowles Commission.¹⁴ In the logic of many disciplines you could make a name for yourself by being skeptical about the previous findings, which gave you an excuse to re-estimate what people had already estimated with your new technique. So there was this intellectual trend to be more and more skeptical about anybody else's claim to identification. Randomized trials became the gold standard for identification.

The point should be made very clearly that there was already a massive and sophisticated randomized control trial effort about social policy in the United States¹⁵ that had existed since the 1970s.¹⁶ I would make the assertion, and then maybe walk it back a bit, that nothing the current randomized trials have done approaches the RAND health insurance experiment¹⁶.

13. See interviews with Michael Kremer (chapter 1) and Jonathan Morduch (chapter 4) for more discussion about these debates and their impact on the movement.

14. The Cowles Commission for Research in Economics, now known as the Cowles Foundation, is a research center focused on using statistical and mathematical models to explain the economy, with a special emphasis on better identification of causality.

15. See interview with Judy Gueron (chapter 5) on this topic, specifically about the lessons learned from using RCTs in social policy in the United States. Also note Gueron's and Howard Rolston's book *Fighting for Reliable Evidence*, which is a thorough history of the use of randomized evaluations in the US, and a discussion along similar lines with Tyler Cowen (chapter 19).

16. The RAND Health Insurance Experiment was one of the first large-scale social policy experiments in the US. The study randomized cost sharing for medical insurance (what we now generally refer to as co-pays and co-insurance). The study found that paying some amount for health services reduced use of those services, but the reduction was equal across highly effective and less effective services. While healthcare improved some health measures among the poorest and sickest patients, overall there was negligible impact on health from consuming less healthcare. Cost sharing also did not change people's risky health behaviors such as smoking. http://www.rand.org/pubs/research_briefs/RB9174.html.

the 1970s in terms of sophistication, in terms of importance, in terms of methodological rigor.

There was this trend that I can sit in a seminar in Cambridge and what instrument you propose, I can concoct a story in which your instrument is wrong. The end game of that is that grad students increasingly feared of getting their dissertations approved because their professors would think of some crazy way in which it might be the case that their instrument didn't produce completely clean identification. The logical consequence of that trend is people resorting to RCTs so they could take the debate about identification off the table. So that was one reason, and it was entirely internal to some subcomponents of the economics discipline.

The second thing¹⁷ is that in aftermath of the cold war, aid was increasingly under attack and increasingly defensive.¹⁸ The World Bank, DFID, USAID, and so forth, had a problem because they had lost a huge part of their political coalition. All of the critiques that programs weren't having the impact they were supposed to have become existentially critical. So all of a sudden, people like Abhijit Banerjee saying things about the World Bank and its lack of impact evaluation had to be responded to in a way that made enormous amounts of money available—because the aid community was politically vulnerable.

The third element is the rise of large-scale philanthropy. Within the development community there have always been development people and policy people. I've always been a development guy. By development, I usually mean national development, large-scale social policies. If you come to me and say I can make that guy better off by putting in place a charitable program to give him food, I'd say, "Of course, you can. Go ahead, but don't call it development." The rise of the MDGs, which were part of the defensive reaction of the post-cold war weakened aid industry, and the rise of people like Gates¹⁸ created a philanthropic interest in precisely the small-scale interventions. After all, you come to people in the World Bank with one of these charity programs and they would say, "You realize this is trivial in terms of what we're trying to do in the national development process, right?" But what's the philanthropist's problem? The philanthropist's problem is that I don't want to bother with national policy, I don't necessarily

This point is expounded on in Pritchett's paper "Can Rich Countries Be Reliable Partners for National Development?"

More specifically meaning the Bill and Melinda Gates Foundation, launched in 2000, and the largest private foundation in the world.

want to work with governments at all. I want to take some cocooned channel of inputs into activities into outputs into outcomes and be sure that actually going to be effective in a way that I can combine some warm glow of giving with some hardheaded analysis.

The only people for which the RCT movement is in fact a tool for the moment is philanthropists. These are guys who are interested in the charity question, not the development question, for the most part. From the charity perspective, there's a nice confluence between the methodological demand for statistical power and of being able to tweak at the individual level. Let me give this person food but not that person, I can give this person a bed but not that person. Because after all, I'm giving something individualized because I'm not trying to affect policy, I'm not trying to affect the government,¹⁹ I'm not trying to affect national development processes. The role of philanthropic concerns in terms of the overall action in developing countries is an important piece.

So you have the confluence of those three things. The methodological stance that there was no clean identification outside of RCTs, the defensive stance of the aid people who had to be seen as responding to these critiques (and the best way to respond was to give money to the people making the critiques), and the rise of philanthropy created this powerful nexus that accounts for the money and effort devoted to RCTs.

TO: Your PDIA²⁰ paper, in some ways is a theory of change involving small changes over time driven by individuals.^[7] The people in the ministries or institutions are running programs and learning things and some number of them are deciding to make changes, some number are trying to tweak things to make them better. There's another critique of the RCT movement that says we should really just be studying institutions. You made points about the "smallness" of RCTs, but you've also made the point that change doesn't happen by remaking the Ministry of Education; it happens from a bunch of people within the Ministry making changes over time.

LP: First of all, PDIA is a theory of change of organizations, not individuals. Its ontological unit is the organization. The whole distinction is separating from the capabilities of organizations that arise from the combination

19. See interview with Frank DeGiovanni (chapter 13) for a discussion of the Foundation's efforts to affect government policy.

20. Problem Driven Iterative Adaptation.

the capacities of individuals and the structures of organizations, not the capacities of individuals alone. We have a role for individuals in the organizational process, but it's a theory of organizational change.

I have a paper called "It Pays to Be Ignorant," which is a positive theory model as to why organizations don't do an RCT.^[8] Imagine that your inputs are coming from an external source. To the extent that the RCT is a tool of an external authorizer, it *has* to be resisted by the organization. Let's say I'm an NGO and I'm getting grants from somebody and I'm competing for grants with lots of other people. And someone comes along and says, "We can do an RCT, and with the RCT you'll discover your true effectiveness." But once you discover your true effectiveness, you're pinned down. You can't game people that it's higher. And you are competing to get grants for, say, microcredit from an organization that might also finance girls' schooling and transparency. But your competitors for those grants aren't pinned down. So the only organization that would take that bet is one that's truly desperate. Part of my ability to persuade the grant maker is by making over-inflated claims of my effectiveness, just like everybody else is doing. As long as that's working, why would I ever agree to be pinned down about my true effectiveness with an RCT? If I'm doing transparency interventions, then I want there to be RCTs of microcredit. Then I can make outrageously huge claims about how effective my transparency initiative is going to be without being pinned down, while the microcredit people are hampered in their persuasion by evidence. The funder might be dubious about my claims and pin me down, but not as much as I'd be risking by doing an RCT. So it pays to be ignorant.

Now in our MeE model, the little "e" is experiential learning.^[9] Our point is that you can do internal RCTs, meaning in the process of implementing your project you can try it multiple ways and judge the relative effectiveness of those ways, but without ever pinning down the outcome effectiveness. Our proposal for PDIA is a little bit hedging that bet, but it's much like the feedback that organizations need. What they need feedback is really about how activities translate into outputs. A lot of organizational dysfunction isn't coming from the fact that they have the wrong causal models from outputs to outcomes; it's that they're not effective as an organization in going from inputs to outputs. In that space you can do lots of things to see what techniques get you more outputs, say, higher loan repayment, without ever putting the core organizational link of outputs to outcomes at risk.

Nobody is against the methodological argument that RCTs produce clean identification of impacts. Everybody understands that. But if you want organizations to do more RCTs, they have to be more aligned with a positive model. If you want to do an RCT to change organizational behavior and you want organizations to do more RCTs, you have to align with the true purposes of the organizations.²¹ Any external use of RCTs by donors, funders, ministries of finance, ministries of planning, whoever, is by its very nature a conflictual thing that is not going to build organizational capability or organizational interest in the learning coming from the RCT.

TO: It's obvious that you think many of the people involved in the RCT movement are very, very smart—and that's why you lament that these folks are spending so much time on RCTs. What's your theory of why these very smart people are "wasting their time"?

LP: If the papers they were writing were being written by public health PhD students at Kansas State, I would think it was an enormously productive use of their resources. But that's not who these people are. The main founders of the movement are all geniuses.

For example, Michael Kremer is an unbelievable genius. If you look at the O-rings paper,²² that could have, and should have, been fundamentally transformational in the way we thought about policy and economic growth and everything else.^[10] But have you heard of it? It disappeared without a trace. I think I have four papers with more citations than the O-ring paper and that's just wrong because that paper is way better. And Esther too. She's

21. See the interview with Rachel Glennerster (chapter 17) on why organizations choose to be better consumers of randomized evaluations, and the interview with Dean Yang where he talks about the skill of designing evaluations that will provide useful operational information to implementing organizations.

22. Based on the space shuttle *Challenger* disaster, caused by a single component, O-rings, failing, this is a theory of how production evolves in firms, and in countries, when the value of an output can be destroyed by a single mistake in the production process. The model can explain large differences between incomes in nations and why firms in less developed countries are primarily small. Marginal Revolution University has a useful explanation of the O-ring theory here: <http://mruniversity.com/courses/development-economics/o-ring-model>.

According to Google Scholar, as of August 2016 Kremer's O-ring paper is his most cited (1549), followed by his deworming RCT paper (1497) and a paper on population growth and technological change (1359)

unbelievably smart. The others I don't know as well but seem equally intellectually endowed and capable.

So the answer is, one, I don't know and, two, it should be obvious that after some time we should have stopped talking about these things because you can't keep talking about them.

They're not bad decisions from a disciplinary point of view. The disciplinary logic was inevitable. And these people took the disciplinary logic toward a big shift in disciplinary approaches. It was a necessary next step. Where they were willfully ignorant in my opinion was on the other two fronts. On the positive theory of policy adoption and on the importance thing.

One reason I often sound very frustrated about this is that it is not as if all of the things I am saying are "lessons learned." I, and many others, were stressing all three points from the very beginning. I was, curiously, the task manager of record at the World Bank of Kremer's textbook experiments in Kenya.^{23[11]} So it isn't like "well, we did experiments for a while and we learned they weren't the panacea (or even very useful medication) like we had hoped." It is more like "we ignored the wise advice of every practitioner about why we were wrong and turns out, they were right and we were wrong after all and we are kind of at square one where we were 10 or 20 years ago on all the big issues."

TO: I'd like to ask about a specific application of RCTs—unconditional cash transfers. This is a case where the question wasn't "what's the exact impact of a cash transfer" but "will people misuse and abuse unconditional cash?" Now we have a bunch of different tests in different contexts with different programs and implementers,²⁴ and they all seem to line up to say that, "no, unconditional cash transfers to poor people aren't 'wasted.'"^[12] I find it hard to believe that you could produce convincing evidence of that

23. In this study, Glewwe, Kremer, and Moulin find, in a randomized trial, that providing textbooks to schoolchildren does not raise average test scores, while raising the scores of the best students. See the interview with Michael Kremer where he comments on this study informing his understanding of the issues with inappropriate curricula in developing world educational contexts.

24. For a useful overview see Blattman and Niehaus's review in *Foreign Affairs*. Notable individual studies include Blattman et al. in Uganda (a comparison of a training program with an unconditional cash grant), Baird et al. in Malawi (comparing a CCT to a UCT), and Haushofer and Shapiro in Kenya (measuring the impact of an unconditional cash grant).

without RCTs. Is that a useful application of the methodology? Are there better, faster, cheaper ways that question could have been answered?

LP: This is a great example as it illustrates all the points.

First, this question was answered and had been answered dozens if not hundreds of times with nonexperimental methods. After all, a pretty fundamental part of economic theory says “money is fungible” and “people maximize welfare” so the incremental impact of “in-kind” and “cash” assistance should be exactly the same. Unless the “in-kind” assistance is a “corner solution”—and even there with resale, it should be like cash with a discount for the transaction to get it into cash. Anyone who could be persuaded by evidence was already persuaded. If people had not been “persuaded” it was due to either organizational interests—NGOs want to deliver in-kind for a variety of reasons, farmers want in-kind delivery of food because it gives them a good price selling to governments—or political interests—they didn’t want to make transfers anyway.

And it is hardly like “unconditional cash transfers” are an innovation. After all, that was the widespread design for many social programs before the RCT-driven fad of “conditional cash transfers.”

Second, you’re saying that people are now convinced whereas they weren’t before. Who are you talking about? What is the evidence that RCTs have been widely more persuasive? Now that people are looking at this, my impression is that being told evidence is “rigorous” doesn’t actually have any additional persuasive power.

Sure, there is a mini-fad in this direction among a few philanthropists, but I have yet to see mainstream NGOs or governments engaged in a wholesale revision of their policies.²⁵ And where this is a debate at scale, in India, for instance, this is a debate that has been ongoing for decades and all the proponents of cash were proponents of cash before any RCTs. That’s because they are mostly economists and hence understood that cash is always superior in any nonpaternalistic approach. What is shifting is the politics of the debate and the evidence of waste and corruption in in-kind delivery is more “persuasive” than evidence that it is not “wasted.”

Third, this illustrates the “small” and “charity” related element of the debate. Evidence shows that if you want large-scale poverty reduction at the national level from high levels of extreme or even “2-dollar-a-day”

25. See the interview with Chris Blattman (chapter 20) where he generally agrees that there has not been much policy change as a result of the studies of unconditional cash transfers.

poverty, you need broad-based economic growth. The East Asian countries²⁶ that eliminated mass poverty in rapid fashion—Korea, Taiwan, Malaysia, Indonesia, China, Vietnam, et cetera—did so through growth. In fact in many, if not most, inequality was stable or rising during the period of rapid poverty reduction. They didn’t do it through unconditional cash versus less effective in-kind transfers. So while some countries like Brazil have made some progress with more effective transfer programs, this was from already being a middle income country with almost no extreme poverty. If I am a philanthropist and want to do charity work, yes, give cash. But again, don’t pretend *that* is “development.”

26. See interview with Tyler Cowen (chapter 19) where he notes that further study of these examples is likely to be one of the most influential fields over the next 30 years.