Macartan Humphreys
2011 Visiting Trudeau Fellow,
University of British Columbia

BIOGRAPHY

Macartan Humphreys is a professor of political science and the director of the Columbia Center for the Study of Development Strategies at Columbia University. His research focuses on the political economy of development, governance, and conflict processes. As a 2011 Visiting Trudeau fellow, Professor Humphreys contributed to academic life and research at the University of British Columbia during the 2011/12 academic year.

Professor Humphreys has published widely in peer-reviewed journals and has co-authored or co-edited two books on ethnic politics and natural resources. He sits on the editorial board of the *American Journal of Political Science* and is a founding member and the current director of the Experiments in Governance and Politics network.

His recent research has pioneered the use of an experimental approach in the study of the political economy of development. Ongoing projects include a field experiment on technological diffusion in Uganda, an experiment on political accountability in Uganda, and a set of experiments on post-conflict development and political participation in Liberia and Congo. Other research has examined ethnic politics in Uganda, the organization of fighting groups in Sierra Leone and Aceh, the political economy of natural-resource management, and the use of information technology to strengthen relations of political accountability. Professor Humphreys has undertaken field research in Chad, Colombia, Democratic Republic of Congo, Gambia, Ghana, Guinea Bissau, Haiti, Indonesia, Liberia, Mali, Mauritania, Morocco, São Tomé e Principe, Senegal, Sierra

Leone, and Uganda. He holds a BA in history and political science from Trinity College Dublin (1994), an MPhil in economics from Oxford (2000), and an MA and PhD in government from Harvard University (1998, 2003).

ABSTRACT

Considerable resources are invested by rich countries into trying to alter social structures in the developing world. But there is little evidence regarding the wisdom or the effectiveness of these interventions. This paper describes a large-scale randomized experiment implemented in East Congo that sought to assess the effects of a major intervention of this form. The study found little evidence of any effects at all, good or bad. These null results, and the research approach used to generate them, raise critical practical and ethical questions for development policy but also for the practice and the communication of research in international development.

LECTURE

"Experimental Research, Development Policy, and Agency Politics in the Congo: Reflections on a Null Result"

Munk School of Global Affairs, University of Toronto

FEBRUARY 6, 2013

Introduction[1]

In early July 2006, I received a call from the International Rescue Committee to discuss the idea of working with it to assess the effects of a large aid program that it was planning to implement in East Congo. I had worked on a similar project in post-conflict Liberia and was growing very interested in this sort of development intervention.[2]

It was easy to see why the research the organization proposed could be important. The type of program in question—community-

1. This paper draws heavily on joint work conducted with Peter van der Windt and Raul Sanchez de la Sierra. Enormous thanks to them both for their leadership in the research and for our many conversations on the issues I discuss here. My thanks to the Pierre Elliott Trudeau Foundation for its generous support and to the University of British Columbia for providing a welcoming and challenging environment while I undertook this research. The full list of people who played critical roles in making this study possible runs to many pages, and I refer readers to the acknowledgements in our paper Humphreys, Sanchez de la Sierra, and van der Windt, "Social and Economic Effects of Tuungane," Working Paper, Columbia University, 2012.

2. James Fearon, Macartan Humphreys, and Jeremy Weinstein, "Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia," *American Economic Review* (P&P) 99, no. 3 (2009), 287-91.

driven reconstruction (CDR), a subclass of community-driven development—is being used by a range of development actors in countries as diverse as Afghanistan, Indonesia, and Liberia. The World Bank estimates that community-driven development programs count for about US$1.3 billion a year in its portfolio alone.

These programs are not just large in terms of scale, they also have grand ambitions. The hallmark of these program is that, rather than engaging with national governments, they allocate development funds directly at the local level—often at the level of villages. In other words, citizens decide how to use the funds. What projects should be supported? Who should benefit? The idea is that making these decisions at a local level is likely to reduce cheating and produce better decisions about how to allocate funding, since the decision makers have every incentive to use the funds as well as possible. This might be called the *efficiency* argument for CDR programming.

CDR programs are also motivated by *intrinsic* or *instrumental* arguments: that it is intrinsically good for people to engage in decisions that affect them,[3] or, more cynically, that it can be politically useful for people to feel that they have a say.

Very often, however, a very different, and much more ambitious, argument is used to justify CDR programs: namely, that CDR is not just effective, but is also *transformative*. This argument holds that CDR does not just leverage the governance gains that arise from bringing decisions down to the local level; it also transforms the nature of governance itself.[4] In post-conflict contexts, this argument

3. The World Bank makes the intrinsic argument, noting that community-driven development "improves not just incomes but also people's empowerment, the lack of which is a form of poverty as well" (World Bank, "IDA at Work—Community-Driven Development: Delivering the Results People Need," 2009; available at http://siteresources.worldbank.org/IDA/Resources/IDA-CDD.pdf).

4. Note my use of the word "governance," which is the term many groups working in this area use; what we are really talking about, however, is politics.

is motivated by the idea that social and political problems are at the heart of development failures, and that to be effective, aid must not only provide material support, but must also seek to induce political change.[5]

The transformative agenda has important implications for how aid takes place. Perhaps the most important implication is that local decision-making structures begin to be seen as the problem rather than the solution. Many CDR programs put pre-existing local institutions to the side and create new local decision-making groups, generally through local elections. Often, the programs place strong external impositions on what the new groups must look like, requiring them to include some populations (women, for example) and to exclude others (traditional leaders, perhaps).

The transformative agenda is intellectually intriguing. Understanding the evolution of political institutions is a holy grail of political science. Political scientists have paid tremendous attention to understanding why some states seem to be set up to provide benefits to their citizens while others appear to be intent on stealing as much as they can, as quickly as possible. According to classic political economy accounts, the key factors are elements such as the size of the middle class, the structure of inequality, and the fiscal demands of the state. Much of this classic work emphasizes structural processes, and generally internal structural processes, that move slowly. Other work emphasizes the role of institutions: set up the decision-making structures correctly, the reasoning goes, and good things follow. These accounts, too, focus on processes that unfold over many decades. This large literature calls into question the feasibility and the wisdom of the transformative agenda.

5. The World Bank summarizes the twin goals, arguing that "community-driven development operations produce two primary types of results: more and better distributed assets, and stronger, more responsive institutions." World Bank, "IDA at Work—Community-Driven Development" (2009).

Outside of the academy, in contrast, there is hope that substantive change can happen relatively quickly and with light-touch interventions. The Congo program had exactly these transformative goals. The program was called Tuungane, Swahili for "Let's Unite." Funded by the Government of the United Kingdom in the amount of US$46 million for the first phase and US$95 million for the second phase, the program aimed to first reorganize existing settlements into new quasi-communities, then set up elections to create project management committees, and, finally, implement development projects in areas selected by the committees in consultation with local populations. The committees would be responsible for overseeing the quality of implementation and for reporting back to the populations; the populations would learn that they could select their leaders democratically, charge them with making decisions, hold them to account, and, at the end of the day, have nice infrastructural development to show for it. As one of the implementers of the CDR program argued, "This program is exciting because it seeks to understand and rebuild the social fabric of communities…It's a program that starts to rebuild trust, it's a grassroots democratization program."[6]

Millions of dollars are being invested in this approach around the world, and some of the results from our work in Liberia seemed to support it. Perhaps there was something to it.

The Study

So I agreed to work with the International Rescue Committee on this project. At the outset, however, I wanted to be sure that if we did it, we could do it credibly. Our study design had many pillars, perhaps the two most important of which were our use of randomization and our reliance on behavioural measures.

6. International Rescue Committee, "In Congo, Learning Democracy and Rebuilding Communities," November 4, 2008, available at www.rescue.org/news/congo-learning-democracy-and-rebuilding-communities-4414.

The idea behind randomization is very simple. Fundamentally, the argument is that because the world is so complex, one needs to do something a little dramatic in order to uncover underlying orders: one needs to inject some disorder. In other words, the complexity of the world comes not from the fact that there is too little order, but from the fact that there are so many orders that interact with each other, magnify each other, and hide each other. Separating one strand of order from another is hard, and generally we fail at it. But true randomness can interrupt these many orders and let individual strands stand out.

For the sort of problem we were looking at—understanding how introducing democratic decision-making institutions would alter local governance structures—one of the biggest challenges for figuring out the effects of the program lay in the normally hidden decision making that would determine where the development organization would operate. More specifically, if the organization were to operate in the most difficult areas of East Congo, skipping over the easier areas, then our analysis in those places would show poorer results than in places where the organization was not operating, possibly causing us to conclude that the program was making things worse. Conversely, some development programs concentrate on better-off places where program managers believe that they can work effectively without putting their staff, or the program, at risk. If this were the case for Tuungane, then our assessment might conclude that the program was having wonderful effects, even if the net effect was negative. Knowing how well an area was faring before the program started would not solve this problem. It is possible, for example, that if all the potential program areas were faring equally well or equally poorly before the program started, program managers might still choose to work in one area rather than another, because they had reason to believe that conditions in that area were likely to get better (or worse).

If, however, program areas were chosen by lottery, then the random selection of program areas would guarantee that there would be no systematic difference between the areas where the program was operating and the areas where it was not (except, of course, for the fact that the program was operating there). This randomness would give us grounds to conclude that any differences between the treatment areas and the control areas were attributable to the program alone.[7]

The arguments for randomization are strong, and we decided to apply it to Tuungane. To do so, Tuungane used public lotteries. Community leaders from different regions met, the names of all of the communities were placed in a hat, and the names of the communities where Tuungane would operate were drawn. In all, 280 communities—each with about 6,000 inhabitants—were selected for treatment, and 280 communities were not.

This left the problem of measurement.

Measurement is always difficult, but it is especially difficult for social outcomes. Classic approaches rely on different types of survey measures. Many clever innovations have improved the quality of data obtained from surveys, but the fear remains that, at the end of the day, people say what they think you want to hear. After all, is it

7. There are various ways that this can go wrong. One is if the lottery has an effect that is distinct from the program. For example, it could happen that being selected to take part in a program convinces a group that they are blessed and that this conviction has effects even if the program has none. For example, people might start making investment decisions on the basis of the fact that they have been selected: positive or negative consequences could arise even if the program never takes place. In medical trials, researchers use placebos to try to counter this type of effect. For many social science interventions, though, a placebo is not possible. It is also possible that areas that are not accepted for treatment are affected by those that are. In that case, too, a simple comparison of outcomes might be misleading. One might, for example, conclude that a program was effective simply because it made non-participants worse off.

not to be expected that after years of being told about the importance of transparency, accountability, and the rest of it, respondents know exactly what answers surveyors wanted to hear? Testimonials routinely provided by practitioner groups have an eerie 1984 quality that we wanted to avoid.[8]

Our study gave us the chance to examine directly this kind of bias (sometimes called "social desirability bias"). When we implemented the endline survey, we asked all of the respondents a straightforward question: "Do you think that elections should be used to appoint people to positions that require expertise?" For half of the respondents, however, we preceded the question with the statement, "Many organizations think that elections are not a good way to choose people for positions that require expertise." For the other half, we preceded the question with, "Many organizations think that elections are always the best way to choose representatives, even for positions that require expertise." Our hypothesis was that if people responded according to their prior convictions, their answers would not vary according to the statement that preceded the question (the "prime"). If, in contrast, respondents tended to give surveyors the answers that they thought the surveyors wanted (or if, more simply, they were easily swayed by arguments without content), then their responses would be very sensitive to the prime.

Some 65 percent of respondents told us that they favoured elections, even when we suggested that organizations maintained that elections were not appropriate for positions requiring expertise. This

8. The International Rescue Committee's website, for example, includes a Q&A with a member of a development committee. "Q: What has your experience been, working with the Tuungane project? A: We've discovered a lot as a community here…Since becoming united through the CDC [Community Development Committee], we've seen that it's important to work together. There is work that one person can't realize, but with the force of everyone, we've realized great things." In "Q&A from Congo: Paving the Way for Women in Leadership" (2008), available at www.rescue.org/news/qa-congo-paving-way-women-leadership-4415.

high number suggests that large numbers of people support elections and are at least minimally willing to argue the point. When we suggested that organizations always favour elections, the number jumped to 84 percent. The effect of the prime was thus close to 20 percentage points. This is an enormous effect, much larger than the real effects that most programs hope to achieve. It suggests that respondents' willingness to please could be large enough to drown out any substantive effects of interest.[9]

So we needed something more reliable, a measure of what people do, not what they say. There has been a huge growth in the use of behavioural measures that seek to do just this. Classic examples involve things like dropping a wallet on the street in different neighbourhoods and seeing when and where the wallet is returned, or dropping stamped envelopes with different names and addresses on them to see which will or will not be picked up and mailed. In our Liberia study, one of our measures assessed how much of a private pot of money an individual was willing to contribute to a community pot. Approaches like this one have the virtue of decreasing the likelihood that subjects will seek to provide the "right" answer. The weakness of such approaches is that knowing what the measures mean is often difficult. If, on average, people in treatment groups are willing to give five cents more of a dollar to a public fund than are people in control groups, is this a big effect or a small one? How much does the difference depend on how the problem was presented to the subjects in the first place, or on other elements that researchers sometimes unwittingly control? The answers to these questions are often not very satisfying, especially for people who want to use findings to inform policy decisions.

9. Interestingly, we found that the bias effect was unrelated to participation in the program. Individuals who were not in the program were just as willing to try to provide the "right" answer and were no less willing to argue for electoral mechanisms.

Our solution was to present communities with a simple collective action problem not unlike one they might face under other circumstances. We introduced a new intervention—RAPID—in both the Tuungane and the non-Tuungane areas. Under the RAPID program, villages with populations of about 800 were given an unconditional community grant of US$1,000. The communities were asked to form a committee to manage the grant (we imposed no requirements regarding the composition of the committee) and to describe how they would use the funding (constraints on admissible uses were very minor).

With this basic structure in place, we hoped to determine whether areas that took part in the Tuungane CDR program engaged differently in RAPID than areas that did not take part in the Tuungane program. Did more people participate in decision making in the Tuungane areas? Was decision making more consultative? Were the ultimate outcomes more or less equitable?

We also wanted to understand how information about development goods spread through the villages. To accomplish this, we introduced a wrinkle: when we introduced the project to communities, we told the whole village that US$900 or more would be made available to it. Once the committee was formed, however, we handed it US$1,000 in private. We were interested in knowing whether information about the extra US$100 would spread through the village.

This structure allowed us to put some simple but tough tests in place. Did taking part in the development program make a difference to these communities, not merely a difference in the language they used, but a difference in the way they made collective decisions—a difference in how politics was done?

What did we find? To our surprise we found nothing. Or nearly nothing. We implemented multiple tests over hundreds of measures, and measure by measure, we found that places with the Tuungane program looked a lot like places without it. We confirmed that there

were elections, that there were meetings, and that there were projects. We also confirmed high levels of beneficiary satisfaction—most people said they liked the CDR project and wanted more of it. But they did not act differently, at least not on the items we looked at. They were no more likely to show up to community discussions about projects, they were no more likely to use voting to make decisions, they were no more likely to spread benefits more broadly, and they were no more likely to have leaders who did not engage in graft. The issue was not that they did not take part or that they did not use elections or that they did not use the money well. Very many of them did all these things. But so did the communities that did not take part in CDR.

In short, it is quite possible that CDR is an effective mechanism for disbursing funds, but we found no evidence that it is transformative.

Burdens

This was a costly study, costly in every sense. Most obviously, it was financially costly. Setting the program up as a randomized intervention meant that it had to cover about twice as large a geographical area as otherwise. Randomization also made heavy demands on the organizations working on the program in terms of data maintenance and the timing of operations. The decision to use behavioural measures added costs, as it required transferring US$1,000 to 560 communities and activating an extensive logistical and security apparatus. Finally, the study imposed costs on the organizations' political capital; these costs grew heavy when the governor of one province became convinced that our research was part of a political campaign against him.

Beyond the finances were heavy human costs. Our team of about 100 enumerators trekked for 18 months through some of the most difficult terrain imaginable. The enumerators spent many months at a time far from their homes and their families. They often had to

walk or push bikes over great distances, and they regularly suffered from malaria, cholera, and other sicknesses.

As in all research, the respondents also bore costs. In this case, thousands of people throughout the region sat for hours answering questions and engaging with our team. The members of the communities that participated in RAPID also engaged with each other, making collective decisions about how to use and distribute scarce resources—a process that can give rise to tensions and conflicts of its own.

The poor security and infrastructure of East Congo made things worse. More than once, staff members delivering payments to villages were taken hostage by armed groups. In a terrible incident unrelated to the project, but chilling for all of us, one of our enumerators was brutally assaulted in her home by a gang of Congolese soldiers, just days before she was meant to go into the field. In another tragic incident, a seven-year-old girl died in a motorcycle crash involving some of the enumerators.

These costs are much higher than those that researchers normally have to worry about, and, in my mind, they placed a lot of responsibility on me and on the research team.

I felt three types of responsibility most strongly. The first was the responsibility not to harm subjects. The second was the responsibility not to make a mess of the research. And the third was the responsibility to make sure that the learning was worth it.

The Responsibility to Do No Harm

The responsibility to do no harm posed numerous challenges. This study was experimental in nature and involved the manipulation of human subjects: we were learning from the fact that some individuals had experienced a program and others had not. The way that we assigned people to treatment groups was orchestrated precisely to allow this learning to take place. In this sense, ours was a randomized experiment. But this experiment and others like it are not

randomized controlled trials for the simple reason that they are not trials. In a classical clinical trial, an intervention is set up to test a drug or treatment: the research question comes first and the intervention follows. In the political economy of development, things are often the other way around. An intervention is decided upon on its own merits; randomization is introduced afterward, to assess the intervention's effects.

In part because of this difference, the ethical standards of experiments in development often seem to fall very far below those used in clinical trials. Consent is often not sought; indeed, subjects frequently do not know that they are part of an experiment or are contributing, via their public actions, to knowledge. Moreover, control subjects do not generally receive a direct benefit. They are not provided with the best-known alternative, and even if the trial is successful, they are often not provided with the treatment.

These differences pose a challenge. On one hand, they mean that the practices of social scientists seem less ethical than those of our colleagues in health. On the other hand, introducing experimental variation is an improvement over more traditional development practice: since interventions happen anyway, it is more ethical to set them up so that we can figure out whether they are actually beneficial (or harmful) than if we do not set them up that way and continue to intervene in the dark.

This said, my view is that insofar as ethics is concerned, we need to improve on two fronts.

First, we need to do better at obtaining consent around manipulation on which this type of research relies. We are aided by the fact that the interventions we examine generally are not harmful; indeed, the rationale for implementing them is the belief that they will prove helpful. Moreover, ex ante lotteries, far from appearing arbitrary, are often seen as fundamentally equitable mechanisms. This was what we found in Congo. We opted for public lotteries so that people

would understand the selection process. People appreciated this because they saw the lotteries as equitable and transparent.

But consent is not just about transparency; it is also about being able to say no. When at the end of the CDR intervention we introduced RAPID, we wanted to be transparent about the fact that RAPID was part of a research project, and we wanted people to be able to opt out. Of course, we did not actually want them to opt out. At the same time, allowing people to opt out of a project worth $1,000 that will generate measures for research purposes is not giving them a real option at all. So we worried that RAPID would wind up coercing people into consenting. In the end, we struck something of a middle ground. When we introduced RAPID, we gave those communities that we invited to participate in RAPID the possibility to opt in, where opting in meant agreeing that data from the project audits would be made available for research purposes. Once a community decided to participate in RAPID and had opted into using audits for research, we requested the right to take more measures—to record features of meetings, to photograph projects, and so on—that communities could refuse without putting their participation in RAPID at risk.

In practice, all of the communities consented to all of the processes. Upon reflection, it is not clear to us that much was gained from the niceties of our consent process. For one thing, for these populations little was at stake; the measurements were unintrusive and anonymity was preserved in any case. For another thing, given the unequal power relations between the villagers and the researchers, it is not clear that villages really felt free to opt out.

The second way that social scientists could work more ethically would be to generate more robust strategies for handling risk. My view now, after my experience with this study, is that the risks associated with studies like ours are often too large for researchers to bear, and that researchers should not bear them. The most important risk

we faced was that harm would come to our enumerators or to our subjects. But sending someone out on a motorbike in Congo or anywhere else risks doing harm. Similarly, providing development aid to a village with deep divisions risks doing harm. So taken literally, the responsibility to do no harm makes no sense. In practice, I think that the principle of "do no harm" really means trying to do more good than harm, or perhaps trying to do much more good than harm.

Under this understanding, our research was entirely consistent with the "do no harm" approach. There are plausible arguments for why development aid may do no good, and why it may sometimes do more harm than good; our study was designed to find out which, in the case of the CDR program, was true. In other words, the problem is not just the presence of risk, but the absence of knowledge.

Even still, our study presented risks of its own. Recall that RAPID imposed very few constraints on how its funds were to be used. We prohibited using the money to buy arms, for example, but we did not take action to ensure that the funding was used equitably. On the contrary: the whole point was to leave communities free to behave inequitably. This approach departed from the standard practice of our partner organizations,[10] but there would have been no point in our putting in place a tool to measure successes if our very study design ensured that there would be no failures. Even so, though, while we were willing to allow failures, we were not willing to force a design that would have allowed risks to escalate. Nor (rightly) would our partners have let us do so.

The team's solution to this question was to share responsibility. While we could apply strong research designs and could bear responsibility for our findings, we could not bear the responsibility for the risks that our designs might produce in the field. Those

10. Somewhat perversely, having high standards to do no harm has meant that in standard programming, projects have not been allowed to fail and conflicts have not been allowed to escalate. If mistakes are a midwife of learning, then this refusal to allow mistakes might explain some of the limited learning.

risks lie within the sphere of responsibility of implementing agencies, which bear them as a part of their daily business and indeed have procedures to minimize them. Agencies are better placed than researchers both to assess the risks and to respond to adverse events. In this case, the agencies wanted to share the responsibility further, and they reached out to their funders in the Government of the United Kingdom for liability protection in the event that project funds were misused.

In sum, while clear lines of support and joint ownership of a strategy do not guarantee that unethical research will not occur, they are nonetheless, in my opinion, an important check on the impulses of scientists who might implement a research design without understanding the risks or without having the capacity to respond should things go wrong.

This said, even if there are clear lines for handling risks and all precautions are taken to minimize them, working in environments like this one still means accepting the risk that something will go wrong. This leaves the biggest question: whether experimental research such as ours should be done at all. In this instance, I find the question unanswerable.

Recall that I mentioned that, during our study, a young girl died in a motorbike accident. Here are some details. At one point during our planning, the question arose as to whether a motorbike could be shared by three people: two enumerators plus a driver. The International Rescue Committee argued against this set-up, which would violate Congolese law. In contrast, one of the local researchers argued strongly in favour, saying that carrying three people on a motorbike was how things were done in the area and was, in fact, the safer alternative, since professional motorbike drivers were better drivers than enumerators. The international organization responded by setting up drivers training for the enumerators. My instinct was to agree with the local researcher, who seemed to be valuing safety over legality. But I knew that I had no particular knowledge to bring to

the table, and so I did not weigh in. It turned out that when the girl was killed in a crash involving our enumerators, the motorbike was carrying three people. The enumerators had decided that even with driver training, it was safer for them to hire a professional driver.

Had I supported the enumerators' position at the outset, I would have felt a very direct responsibility for the death of the girl. But I still did not feel right about not having supported their request to use a driver. Either way, their decision to hire a driver might have been the right one. This particular question about how best to minimize risks seems almost impossible to answer; the truth is that there were risks either way. The greater problem concerns the risks that come with doing anything at scale in this kind of environment. Given environmental risks such as this one, should research at this scale be done at all?

Whatever the answer, one thing is clear: if research like ours is done, the learning had better be worth it.

The Responsibility Not to Mess Up

Researchers generally do not want to make mistakes. In fact, we organize a lot of our work around catching one another's mistakes. But as a group, we are probably more tolerant of mistakes than many. My grandfather used to say that the person who never made a mistake, never made anything. I encourage my students to do projects that involve reasonable chances of getting things wrong, but to learn from doing them. Even so, I was particularly worried about making mistakes in this study, first because so many people were interested in it and were following it, but second, and more important, because the findings might matter.[11]

11. Or more precisely, the way in which they might matter was more obvious. A lot of research matters but often in very diffuse ways. In this case we could expect a fairly direct relationship between research results and decisions regarding the use of future development aid.

We saw the null results coming from a long way off. The way we had set up our study allowed us to analyze the data as it came in. We performed our first analyses with hardly any data (in fact, we performed them with fake data) and so did not expect to see any patterns. As more data came in, we expected to see things settle down and patterns to strengthen. But they did not. All of our interim reports painted a similar picture.

Coincidentally, just as we started our analysis, the Oscillation Project with Emulsion-tRacking Apparatus (OPERA) research team at the European Organization for Nuclear Research reported finding neutrinos that had arrived 60 nanoseconds faster than they would have if they had travelled at the speed of light. Theirs was the first recorded incidence of particles moving faster than light. I felt a tremendous sympathy for the scientists making this report. On the face of it, they were reporting what could be a revolutionary finding. But they also knew that they were probably wrong. Their finding was inconsistent with all existing theory and evidence; it looked to all the world as though they had made a mistake. So they checked their instruments and their results many times before making the announcement. They could not find any mistakes, so they presented their results and shared their data and continued to investigate. Sure enough, as others replicated their work and as investigations continued, it became clear that mistakes had been made. The problem was a loose cable. The particles had never travelled faster than the speed of light; things were as we had always thought they were. As the OPERA team made its announcement to this effect, it was clear that it was a formidable research group working with high standards of integrity. But while one could admire them, it was hard to envy them as they made their announcement.

Back in the more prosaic world of trying to figure out the impacts of CDR programs, our first goal was to make sure that we had not made any mistakes. In practice, this took the form of thousands

of data checks and robustness checks of various types. Checking our code, checking how things would look with different analyses, and so on. But things did not budge. The zeros stayed at zero.

Of course, the usefulness of all this checking and rechecking depended on the measures we were using. Were we using the right measures?

Here we found some solace from a strategy that we had adopted early on, before the results started coming in.

It turns out that statistics lie. Or at least they can be made to lie. In fact, some of us have mastered the art of making the data lie without even knowing we are doing it. Often when researchers meet data, they do not go straight to the business of analysis. First there is a get-to-know-you period when they try to get a feel for the data, how it is constructed, what its strong and weak points are. This is sometimes followed by a listening period when the researchers attempt to find out what the data is trying to say. They let the data speak. Next, they have a conversation, following the interesting patterns until they have a story that they can take away with them. Researchers can almost always do this. In the words of economist Ronald Coase, if you torture the data long enough, Nature will confess.

The only problem is that there is a good chance that Nature will make false confessions. When you follow the interesting patterns, you can end up settling on findings that are entirely spurious. This is called data fishing: the practice of extracting a finding from a pool of data and displaying it to the world without the full statistical context. In general, with a large data set like the one we were working with, you can poke about and select patterns that tell a rosy or a tragic story about the intervention. To a large extent, it would be your call.

The pressure to fish comes from many sides. It comes partly from researchers who find positive results more interesting than negative ones. It also comes from peers. The normal procedure in statistical analysis is to formulate a test, present the results, and gather feedback in the form of suggestions that the researcher try

this or that. Indeed, review and publication processes are organized around constructing or reconstructing tests after seeing the results from prior analysis. Simple analyses of the results published in political science journals conclude that without a shadow of a doubt, our discipline is deeply involved in data fishing. The practice is rampant.

In our case, we were concerned that, depending on what we found, various stakeholders might come back with new ideas for how to conduct the analysis. Some of these ideas would be bound to produce some results, even if they were entirely spurious. But at that stage, it would be very hard to know what to believe. The first results or the new results? How to weigh the two?

We resolved this dilemma by doing something that researchers often think about but almost never do. Before analyzing the real data, we specified our statistical tests down to the finest detail and wrote up an entire mock report based on simulated data. We then shared this report with the project implementers and the UK government so that we could nail down the tests before running them. Our thinking was that if anyone had an idea about better things to measure or better ways to conduct the analysis, he or she could say so up front, not after having seen the results. In effect, we practised a form of research registration—a practice that is now standard in medicine but that has not yet gained traction in the social sciences.

Writing and registering a mock report gave us political cover. It also gave us confidence that the areas where we expected to see change were the same as those where the implementers and the funders expected to see change. This shared ownership over the standards of interpretation helped lessen the burden of being wrong. It also revealed a pattern that I have since seen repeated: namely, people are much better at reacting to existing results than they are at thinking about future ones. Ex ante, people found it hard to say what they expected to happen and what the right indicators of success would be. The creative energy really starts flowing only once the results are in.

A final step we took to lessen the burden of being wrong was to embrace transparency. We made all of our protocols and instruments publicly available before conducting any analyses, and we made the core data available as soon as we had finished running the tests. Our view was that if we had made mistakes, then the sooner they were discovered, the better.

The Responsibility to Make the Research Matter

Was it worth it?

Whether our research was worth it depends in large part on what is done with it next. The normal procedure with research of this form is to write up a lay-language report to share with partners, then to write an academic piece to share with colleagues. The normal response for both agencies and journals is to put null results to the side.

We have been trying to do a lot more than that, to make sure that our results are read and absorbed. As of the time of writing, we have presented our findings to development organizations and to the Government of the United Kingdom in London, Nairobi, Kinshasa, and elsewhere. The study has been covered in the *Financial Times* and has been featured in blogs of the World Bank and the United Kingdom's Department for International Development.

But it can be tempting to push things too far. In thinking about how to present and communicate our findings, we have struggled with four challenges.

The first challenge was not to exaggerate. Researchers might like clear positive results best, but they still prefer clear null results to ambiguous ones. But the truth is that there are always ambiguities. In our case, the economic outcomes of the project are especially ambiguous. We found no evidence that the projects we evaluated had economic effects. But I do not think that our study really provides an argument against infrastructure investments. We deliberately chose to take measurements at a time when governance effects

might be strong but economic effects might be weak. The effects from infrastructure projects such as schools take time to manifest, and it is not surprising that we did not find effects so soon after the project had been implemented.

Our second challenge was how to handle the politics of agency. When things go wrong, people often ask who is to blame. In this case, it is not clear that anyone is to blame. In fact, one of the oddest things about the results was how few people they surprised. High turnover in the agencies has meant that few of those who have read our results were present when the project was designed. In other words, the findings were important for the institutions involved, but less critical for the individuals who work in those institutions. In most cases, staff members had done their job well. The project implementers had conducted extraordinary work under tough conditions, and the designers had adopted models that had been promoted by numerous agencies and organizations (albeit without much proof). Furthermore, almost uniquely in the industry, the agencies had gone out on a limb to test the model they were employing. While they are not now actively publicizing the results, they are working hard to absorb them and to figure out whether and how to continue doing CDR. The systemic nature of the problems with the model will, I hope, make it easier for our findings to make a difference.

Our third challenge was to put our results in context. Ours is just one study, albeit a large one. Major policy decisions should depend on an accumulation of knowledge. So despite all our investment and work, our results should still be seen as a single data point. If other studies show very different results in other parts of the world, then it could well be that the effects we found can be attributed to Congo, or to the Tuungane program, or to our research design, but not to the CDR model as a whole.

Our final challenge was speculation. As soon as you hear that a program of this magnitude has failed, you ask why. A really satisfying explanation would pinpoint failures in design and in

implementation and would describe how to do things better next time. I have lots of ideas for why the program did not work. I think that excluding local leaders was probably a bad idea; I think that the investments were too small, and that communities' projects should have been allowed to fail. I also think that changing politics probably requires changing structures. To many of the would-be beneficiaries, the whole enterprise probably seemed odd: outsiders coming in to introduce consultative institutions using a design that was not based on consultations with the local population and that ignored existing mechanisms for local consultations.

But I also know that these explanations are speculative, and it seems particularly inappropriate to push speculations for why a program was unsuccessful on the back of the five years of research that it took to establish that the program was unsuccessful. Answering those questions, rather than speculating on them, is the stuff of future work.

Conclusions

Let me close by emphasizing two themes.

The first theme is about the production of knowledge. Marx argued that knowledge is socially produced, even if it is privately appropriated. Most of the ideas we have and the concepts we work with have originated with others; most of our innovations are on the margins. Just as importantly, many of the effects of knowledge production are social as well. In our study, joint ownership with the program implementers was critical. We had to make sure that our voice remained independent and we had to be able to disagree, but we also had to coordinate our efforts in order to implement our research responsibly and to make sure that it would be meaningful and could have an impact.

It is striking, however, that while knowledge production is social, scholarly incentives often do not reflect this truth, at least in social science. Scholars compete to claim ownership of ideas; they hoard

their data, and they do not make their instruments and protocols public. Our system values innovation, not verification or replication. But if we want our research to make a difference in the world—and not just a difference to our careers—this is the wrong way to operate. We need to recognize the social character of research, to engage in building common agendas and joint designs, and to be willing to be wrong.

The second theme I want to emphasize relates to development practice. Perhaps the greatest oddity about this study is how quickly beliefs changed after our results came in. Before we gathered any measures, we surveyed practitioners and researchers, asking them what they thought we would find. We learned that respondents expected strong or very strong effects. It is true that some thought that for particular items, the effects would be weak, but on average, respondents expected clear positive effects. Once the results came in, however, the first reactions of most researchers and many practitioners were, Why are you surprised? How could you possibly expect that a program like this one would have substantial effects? Looking at it in hindsight, it does seem a little odd. But while there were early questions over the wisdom of the intervention, the possibility that the intervention would generate effects did not seem so odd before the results came in. On reflection, I think we fell prey to what might be called the "TED talk" fallacy: the mistake of constantly focusing on the most optimistic scenarios and thinking that just because big effects often result from small actions, small actions will probably lead to big effects. In truth, most small actions probably have small effects, if they have any effects at all. The lesson, then, is to try to think in a more ex post way, ex ante.