Should the Randomistas Rule?

MARTIN RAVALLION

Over the last five years or so, an influential group of academic economists—the "randomistas"—has been advocating social experiments as the main tool for studying development effectiveness. In a social experiment some units are randomly assigned an intervention while the rest form the control group, and one compares the average outcomes of the two. The randomistas see this as the only clean way of identifying impact, as it appears to avoid untestable identifying assumptions based on economic theory or other sources. They view non-experimental methods as (by and large) unscientific and best avoided.

The randomistas, with MIT's Poverty Action Lab at the forefront, are gaining influence in the development community. Researchers are turning down opportunities to evaluate public programs when randomization is not feasible. Doctoral students are searching for something to randomize. Philanthropic agencies are sometimes unwilling to fund non-experimental evaluations. Even the World Bank is responding, having been criticized by the randomistas for not doing enough social experiments on its lending operations; the largest fund (to date) devoted to impact evaluations at the Bank gives an explicit preference for randomized designs.

Has this gone too far? To assess the randomistas' case one must first understand the problems in the market for knowledge about development effectiveness. One must then look closely at both the strengths and weaknesses of social experiments in addressing those problems.

THE DISTORTED MARKET FOR KNOWLEDGE ABOUT DEVELOPMENT EFFECTIVENESS

The portfolio of evaluations is probably biased towards things that work reasonably well; the managers and agencies responsible for weak projects try to avoid evaluations. Publication biases do not help either: it is more difficult to publish "no impact" findings. Also short-term impacts get more attention than impacts emerging beyond the project's disbursement period; indeed, evaluations of the long-term impacts of development projects are rare. And we know more about some types of interventions (notably transfers and other social sector programs) than others (such as physical infrastructure).

Martin Ravallion is director of the World Bank's research department. These are the views of the author and should not be attributed to the World Bank.

Decisions about what gets evaluated, and how much is spent on doing so, are typically taken by a rather narrow group of direct stakeholders, funding or managing the individual projects. Yet a large share of the gains from new knowledge is shared with external parties, including future projects. This combination of decentralized decision making about project evaluation with externalities in knowledge generation creates imbalances between what we know and what we want to know.

What then happens if we let the randomistas rule? Plainly randomization is only feasible for a non-random subset of the interventions and settings relevant to development. For example, it is rarely feasible to randomize the location of infrastructure projects and related programs, which are core activities in any poor country's development strategy. The very idea of randomized assignment is antithetical to most development programs, which typically aim to reach certain types of people or places. Governments will (hopefully) be able to do better in reaching poor people than would a random assignment. Randomization is also better suited to relatively simple projects, with easily identified "participants" and "non-participants."

Social experiments also raise ethical concerns, with context-specific political sensitivities. The concerns stem from the fact that some of those to which a program is randomly assigned will almost certainly not need it, while some in the control group will. (Non-experimental methods can raise similar problems.) The randomization is often done "conditional on observables;" in other words, one first selects participants based on the available data, and only then randomly assigns the intervention. (For example, one might only randomize among those who appear to be "poor" based on some criteria.) The ethical problem remains, however, given that the evaluator can observe only a subset of what is seen on the ground (which is, after all, a reason for randomizing in the first place). At local level, there will typically be more information—again revealing that the program is being assigned to some who do not need it, and withheld from others who do. The development randomistas have not said much about these issues and appear to dismiss them (although these issues have been very important in bio-medical experimentation).

The upshot of all this is that the feasibility of doing social experiments varies from one setting to another, and this need not fit well with our knowledge gaps.

The prospect for addressing our pressing knowledge needs is also limited by the fact that social experiments focus on just two parameters among many of interest. The first is the average impact of an intervention on the units that are given the opportunity to take it up; this is called the "intent-to-treat" (ITT) parameter. The second is the average impact on those who receive it; this is the so-called "average treatment effect on the treated" (ATET). (Thinking of development as a "treatment" is highly unfortunate, but the word has stuck.)

I have rarely found that policy makers care only about these two parameters. They also want to answer questions like: Does the intervention work the way it was intended? What types of people gain, and what types lose? What proportion of the participants benefit? What happens when the program is scaled up? How might it be designed differently to enhance impact?

From the point of view of development policy-making, the main problem in the randomistas agenda is that they have put their preferred method ahead of the questions that

emerge from our knowledge gaps. Indeed, in some respects (such as the sectoral allocation of research) the randomistas' success may have made things worse. The risk is that we end up with lots of social experiments that provide evidence on just one or two parameters for a rather narrow set of assigned interventions and settings. The knowledge gaps persist and even widen.

In reality a rich set of non-experimental tools is available for addressing the issues of concern to policy makers. These methods invariably require assumptions, notably about the behavior of key stakeholders, including participants. The randomistas are critical of those assumptions. Their position appears to be that the knowledge gaps that can't be addressed by social experiments are essentially lost causes for scientific inquiry. Yet on closer inspection one finds that the randomistas are making assumptions that are no less questionable in principle than those they don't like in non-experimental work.

THE LIMITED VALIDITY OF SOCIAL EXPERIMENTS

The claimed strength of a social experiment, relatively to non-experimental methods, is that few assumptions are required to establish its internal validity in identifying a project's impact. The identification is not assumption-free. People are (typically and thankfully) free agents who make purposive choices about whether or not they should take up an assigned intervention. As is well understood by the randomistas, one needs to correct for such selective compliance to get ATET right in expectation. The standard solution is to use the randomized assignment as the "instrumental variable" (IV) for program participation; the IV aims to identify the exogenous component of the variance in program placement. The randomized assignment is assumed to only affect outcomes through treatment status (the "exclusion restriction").

There is another, more troubling, assumption just under the surface. Inferences are muddied by the presence of some latent factor—unobserved by the evaluator but known to the participant—that influences the individual-specific impact of the program in question. (This is what Jim Heckman and co-authors term "essential heterogeneity".) Then the standard IV method for identifying ATET is no longer valid, even when the IV is a randomized assignment. There are ways of correcting for this problem, though current practice is lagging greatly. Most social experiments in practice make the implicit and implausible assumption that the program has the same impact for everyone.

While internal validity for ITT and ATET is the claimed strength of an experiment, its acknowledged weakness is external validity—the ability to learn from an evaluation about how the specific intervention will work in other settings and at larger scales. The randomistas see themselves as the guys with the lab coats—the scientists—while other types, the "policy analysts," worry about things like external validity. Yet it is hard to argue that external validity is less important than internal validity when trying to enhance development effectiveness against poverty; nor is external validity any less legitimate as a topic for scientific inquiry.

Heterogeneity in impacts also comes to the fore in thinking about the external validity of social experiments. Inferences for other settings, or scaling up in the same setting, based on the results of a randomized trial can be way off the mark. This can stem from heterogeneity of participants or from contextual factors.

For example, an intervention run by an NGO might be totally different when applied at scale by government officials facing different incentives. Understanding the institutional-implementation factors that make the same program successful in one place but not another is an important but under-researched issue.

When the heterogeneity is in observables, the randomistas argue for replication across different types of participants and settings, to map out all the possibilities. It is far from evident that this is a viable research strategy. The dimensionality of the problem could well be prohibitive. Nor are individual researchers likely to be willing to do near endless replications of the same method for the same program, which are unlikely to have good prospects for publication.

Essential heterogeneity poses an even bigger problem for external validity. A social experiment randomly mixes low-impact people (for whom the program brings little, or even negative, benefit) with high-impact people. Naturally the (non-random) scaled up program will tend to have higher representation from the high-impact types. Given this purposive selection, the national program is fundamentally different to the randomized trial, which may well contain rather little useful information for scaling up. This is an instance of a more general point that many things can change—inputs and even the program itself—when a pilot is scaled up.

A further problem that muddies the waters of inference in practice (for both internal and external validity) is "spillover effects." Recall that we are talking about assigned programs, such that it is meaningful to talk about a set of "non-participants." The further assumption made by the randomistas is that non-participants are unaffected by the program. This is known to be implausible in many applications, especially when we are trying to assess impacts beyond the short term. The extent of this problem depends on the setting, not whether the evaluation was randomized or not.

Spillover effects are pervasive in development applications. Take the example of probably the most common form of an aid-financed development project, namely a poor-area development program, which targets aid to poor counties or villages. Spillovers can stem from the fact that poor areas trade in essentially the same markets as non-poor areas, and that there can be mobility in and out of a poor area. A further (less well recognized) problem is that the targeted poor areas exist within a multi-level government structure. There is a central government, of course, and this is the government the aid donor mainly deals with. But local governments operate in both the targeted villages and the non-participant villages. When the aid agency picks certain villages, the local government above the village level will probably divert its own resources elsewhere; doing otherwise would be neither efficient nor fair. My own research has found credible (non-experimental) evidence that this happens in poor-area development programs in China. Where does some of the spillover go? Yes, it ends up in the comparison villages. Even with randomized assignment—though that is unlikely to be feasible for a poor-area development program—one will underestimate the impact of the aid on the targeted villages.

None of this means that an IV from randomization (either of the intervention or of some key determinant of its placement) is useless; far from it. This relatively clean IV can help throw light on a potentially wide range of structural parameters. This can help us un-

derstand a program's impacts and facilitate interesting simulations of alternative policy designs. That brings us back to the same sort of assumption-driven, theory-based, empirics that the randomistas want to avoid. That was, arguably, their biggest "false lead" for development research.

CONCLUSION

The emphasis that researchers are now giving to obtaining better knowledge about development effectiveness is welcome. Randomization is one of the tools that can help. However, the important task of investigating what works and what does not in the fight against poverty cannot be monopolized by one method.

Letters commenting on this piece or others may be submitted at http://www.bepress.com/cgi/submit.cgi?context=ev.

REFERENCES AND FURTHER READING

Banerjee, Abhijit (2007) *Making Aid Work*, Cambridge: MIT Press.

Chen, Shaohua, Ren Mu and Martin Ravallion (2008) "Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China," World Bank Policy Research Working Paper No. 4084. March. Available at: http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2008/03/03/000158349_20080303131839/Rendered/PDF/wps4084.pdf.

Duflo, Esther and Michael Kremer (2005) "Use of Randomization in the Evaluation of Development Effectiveness," from George Pitman, Osvaldo Feinstein and Gregory Ingram, ed., *Evaluating Development Effectiveness*, New Brunwick: Transaction Publishers.

Heckman James and Jeffrey Smith (1995) "Assessing the Case for Social Experiments," Journal of Economic Perspectives, 9(2): 85–110.

Heckman, James, Serio Urzua and Edward Vytlacil (2006) "Understanding Instrumental Variables in Models with Essential Heterogeneity," Review of Economics and Statistics, 88(3): 389–432.

Keane, Michael (2005) "Structural vs. Atheoretic Approaches to Econometrics," Keynote Address at the Duke Conference on Structural Models in Labor, Aging and Health. September 17–19. Available at: http://www.business.uts.edu.au/finance/staff/MichaelK/JE_Keynote_6.pdf.

Moffitt, Robert (2006) "Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective," from Barbara Schneider and Sarah-Kathryn McDonald, ed., *Scale-Up in Education: Ideas in Principle*, Lanham: Rowman and Littlefield.

Ravallion, Martin (2008) "Evaluating Anti-Poverty Programs," edited by Paul Schultz and John Strauss, Handbook of Development Economics Volume 4, Amsterdam: North-Holland.

Rodrik, Dani (2008) "The New Development Economics: We Shall Experiment, But How Shall we Learn?" Brookings Global Economy and Development Conference. May. Available at: http://ksghome.harvard.edu/~drodrik/The%20New%20Development%20Economics.pdf.