



Program Evaluation: Are we ready for RCTs?

By Jodi Nelson, Director of Research & Evaluation, International Rescue Committee

It is clear that a new wind is blowing in the discussion about how best to evaluate international aid. Three back-to-back events in Washington, DC recently attracted the attention of bilateral agencies, government ministries and private donors: December's announcement of the first director for the International Initiative for Impact Evaluation (or 3IE) was followed by a conference held by the World Bank, entitled "Making Smart Policy: Using Impact Evaluation for Policy Making," and the convening of members of the Network of Networks on Impact Evaluation (NONIE). These meetings shared a focus on rigorous impact evaluation: "analyses that measure the net change in outcomes for a particular group of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question that is being investigated and to the specific context," to quote 3IE's founding document of March 2007.

For some, the phrase above hints at an ominous turn toward *randomized control trials* (RCTs) – the epitome of experimental research and an investigative approach assumed by many to be largely inappropriate for the contexts in which NGOs work. Online discussions suggest concern about a donor-driven evaluation agenda and the overly scientific measurement of tangible results. The worry likely stems from two views. First, that assigning aid programs randomly is unethical because it contradicts principles about serving those in need, the most vulnerable or excluded. And second, that the use of "rigorous" approaches contradicts all we have learned about the importance of evaluation as a means of empowerment, not just measurement.

The International Rescue Committee (IRC) has in motion four "randomized impact evaluations." In the hope of beginning to demystify this new catch phrase, we would like to share five of the many lessons we have learned

through ongoing efforts in Afghanistan, Burundi, the Democratic Republic of Congo and Liberia.

1. Random assignment is the least challenging component of these evaluations.

Where there were “true” controls (i.e., a group of people that will not get the program), people appreciated the transparent allocation of scarce resources. This confirms how intuitive it is to make and share decisions about where to work based on principles of fairness and independence. We have also learned that randomization means randomly assigning any aspect of the program, not only the full package of whatever comprehensive program is implemented. IRC’s experience in Burundi confirms that you can modify program design and randomly select which group gets which variation without using a pure control.

2. The quality of population data that random assignment requires is one of the biggest challenges to evaluating programs in post conflict contexts.

In the Democratic Republic of Congo and Liberia, we needed the number and names of the communities in which IRC would possibly work, the number of villages within each larger unit and the number of households within each village. It became very clear that population data is an under-resourced public good for these countries and for agencies seeking to support their development.

3. The quality of academic evaluators with whom IRC has worked to date has been impressive, but their supply appears low.

Their flexibility, commitment and growing experience in both NGOs and the contexts in which we work are a strong indication of the deep flaws in the stereotype of “ivory tower” scholars out of touch with the “real world.” But there are still relatively few of these individuals given the number of evaluations we would need to conduct to genuinely improve our knowledge of what works. Experience suggests that there are low barriers to entry for evaluators; a doctoral candidate doing his first such evaluation is assessing a flagship community-driven reconstruction program in Afghanistan. If impact evaluation continues to gain popularity, it will be important to know who will determine criteria for both evaluators and evaluation design.

4. The conventional wisdom on impact evaluations is that it would be counter-productive to evaluate all our work using random assignment.

Some programs simply cannot be evaluated using random assignment, and others do not need to be. The more important lesson is that the evaluation design should match the context, program and the questions we want answered. When we ask what change our work produces – i.e., the net difference we make for people – randomized evaluation is a good approach to consider.

This conventional wisdom also says we should use impact evaluation to study strategic priorities that represent substantive gaps in our knowledge. Doing what is called “variation in treatment” – evaluating the relative effectiveness of multiple approaches to achieve the same objective – highlighted for IRC how difficult it can be to identify these priorities and, more specifically, the questions we need to answer to improve outcomes for people. It is certain that doing so requires greater collaboration and consultation so that we improve our chances for building strategic, cumulative knowledge.

5. A “good” randomized impact evaluation requires not just a survey of treatment and control before and after the program, but good quality monitoring (using a variety of research methods depending on what you are trying to understand). Without this layer of data collection, we run the risk of claiming that we know something “works” but not really understanding why; or, likewise, thinking that it does not work and not understanding why. This is particularly true for social programs in which qualitative variables tend to be important and for programs implemented in difficult contexts where logistics and capacity can be critical.

There are many more lessons IRC continues to learn through ongoing experience. Perhaps the most important is that doing “real” impact evaluation is hard. This seems to make more sense than the assumptions underlying a good amount of current aid evaluation: that explanatory analysis (that x leads to y) is easy to generate; and that any evaluation can answer any question, regardless of its design and choice of data collection methods. Indeed, governments in Central and South America (and increasingly in Africa as well) are using random assignment to evaluate their policies. At the very least, NGOs should follow suit and be informed about what it means and why it is distinct from more conventional practice. Even if it ends up being a fad, the discourse on “real” evaluation can remind us not to take evaluation lightly. Different evaluation designs do not serve the same purpose or answer the same questions.

If it is not a fad but the beginning of a lasting trend, we should be careful not to miss the opportunity to inform what could turn into a very complicated donor requirement in the near future. ●●○