Making Impact Evaluations More Useful

Martin Ravallion

Development Research Group (DECRG), World Bank

In the absence of strong institutional support we under-invest in evaluations

- Development is a learning process, in which future practitioners benefit from current research.
  - Evaluative research is (in part) a public good, in that the benefits spillover to other development projects.
  - Larger externalities for some types of evaluation (first of its kind; "clones" expected; more innovative)
- But current individual projects often hold the purse strings.
- The project manager will typically not take account of the external benefits when deciding how much to spend on evaluative research.

=> under-investment in evaluations, without strong support from outside the project.

However, the problem goes deeper

Evaluations are often not as relevant for practitioners as they could be.

- Classic concern is with internal validity for mean treatment effect on the treated for an assigned program with no spillover effects.
- And internal validity is mainly judged by how well one has dealt with selection bias due to unobservables.

This approach has severely constrained the relevance of impact evaluation to development policy making.

Ten steps to more policy-relevant impact evaluations

1: Start with a policy-relevant question and be eclectic on methods

- Policy relevant evaluation must start with interesting and important questions.
- But instead many evaluators start with a preferred method and look for questions that can be addressed with that method.
- By constraining evaluative research to situations in which one favorite method is feasible, research may exclude many of the most important and pressing development questions.

Standard methods often don't address all the policy-relevant questions

- *What is the relevant counterfactual?*
  - "Do nothing": that is rare; but how to identify more relevant CF?
- *What are the relevant parameters to estimate?*
  - Mean vs. poverty (marginal distribution)
  - Average vs. marginal impact
  - Joint distribution of Y^T and Y^C , esp., if some participants are worse off: ATE only gives net gain for participants.
  - Policy effects vs. structural parameters.
- *What are the lessons for scaling up?*
- *Why did the program have (or not have) impact?*

2. Take the ethical objections and political sensitivities seriously; policy makers do!

- Pilots (using NGOs) can often get away with methods not acceptable to governments accountable to voters.
- Deliberately denying a program to those who need it and providing the program to some who do not.
  - Yes, too few resources to go around. *But is randomization the fairest solution to limited resources?*
  - *What does one condition on in conditional randomizations?*
- Intention-to-treat helps alleviate these concerns
  => randomize assignment, but free to not participate
- But even then, the “randomized out” group may include people in great need.

The information available to the evaluator (for conditioning impacts) is a partial subset of the information available “on the ground” (incl. voters)

3. Taking a comprehensive approach to the sources of bias

- Two sources of selection bias: observables and unobservables (to the evaluator) i.e., participants have latent attributes that yield higher/lower outcomes
- Some economists have become obsessed with the latter bias, while ignoring enumerable other biases/problems.
    - Less than ideal methods of controlling for observable heterogeneity including *ad hoc* models of outcomes.
    - Evidence that we have given too little attention to the problem of selection bias based on observables.
    - Arbitrary preferences for one conditional independence assumption (exclusion restrictions) over another (conditional exogeneity of placement)
    - Cannot scientifically judge appropriate assumptions/methods independently of program, setting and data.

4. Do a better job on spillover effects

- *Are there hidden impacts for non-participants?*
- Spillover effects can stem from:
  - Markets
  - Behavior of participants/non-participants
  - Behavior of intervening agents (governmental/NGO)

Example 1: Employment Guarantee Scheme

- assigned program, but no valid comparison group.

Example 2: Southwest China Poverty Reduction Program

- displacement of local government spending in treatment villages

=> substantial underestimation of impact

5. Take a sectoral approach, recognizing fungibility/flypaper effects

- Fungibility
  - You are not in fact evaluating what the extra public resources (incl. aid) actually financed.
  - So your evaluation may be deceptive about the true impact of those resources.

- Flypaper effects
  - Impacts may well be found largely within the “sector”.
  - Need for a broad sectoral approach

6. Fully explore impact heterogeneity

- Impacts will vary with participant characteristics (including those not observed by the evaluator) and context.
- Participant heterogeneity
  - Interaction effects
  - Also essential heterogeneity + participant responses (Heckman-Urzua-Vytlacil)
  - Implications for:
    - evaluation methods (local instrumental variables estimator)
    - project design and even whether the project can have any impact. (Example from China's SWPRP.)
    - external validity (generalizability) =>
- Contextual heterogeneity
  - *"In certain settings anything works, in others everything fails"*
  - Local institutional factors in development impact
    - Example of Bangladesh's Food-for-Education program
    - Same program works well in one village, but fails hopelessly nearby

7. Take "scaling up" seriously

With scaling up:

- Inputs change:
  - Entry effects: nature and composition of those who "sign up" changes with scale.
  - Migration responses.
- Intervention changes:
  - Resources effects on the intervention
- Outcome changes
  - Lags in outcome responses
  - Market responses (partial equilibrium assumptions are fine for a pilot but not when scaled up)
  - Social effects/political economy effects; early vs. late capture.

But there has been little work on external validity and scaling up.

Examples of external invalidity: *Scaling up from randomized pilots*

- The people normally attracted to a program do not have the same characteristics as those randomly assigned + impacts vary with those characteristics

=>"randomization bias" (Heckman & Smith)

- We have evaluated a different program to the one that actually gets implemented nationally!

8. Understand what determines impact

- Replication across differing contexts
  - Example of Bangladesh's FFE:
    - inequality etc within village => outcomes of program
    - Implications for sample design => trade off between precision of overall impact estimates and ability to explain impact heterogeneity
- Intermediate indicators
  - Example of China's SWPRP
    - Small impact on consumption poverty
    - But large share of gains were saved
- Qualitative research/mixed methods
  - Test the assumptions ("theory-based evaluation")
  - But poor substitute for assessing impacts on final outcome

9. Don't reject theory and structural modeling

- Standard evaluations are "black boxes": they give policy effects in specific settings but not structural parameters (as relevant to other settings).
- Structural methods allow us to simulate changes in program design or setting.
- However, assumptions are needed. (The same is true for black box social experiments.) That is the role of theory.

- *PROGRESA* example (Attanasio et al.; Todd & Wolpin)
    - Modeling schooling choices using randomized assignment for identification
    - Budget-neutral switch from primary to secondary subsidy would increase impact

10. Develop capabilities for evaluation within developing countries

- Strive for a culture of evidence-based evaluation practice.
  - China example: “Seeking truth from fact” + role of research

- Evaluation is a natural addition to the roles of the government’s sample survey unit.
  - Independence/integrity should already be in place.
  - Connectivity to other public agencies may be a bigger problem.

- Sometimes a private evaluation capability will be required.