Eighth Annual Meeting of the Impact Evaluation Network, IADB, 2014

How can evaluations better inform public spending decisions?

Martin Ravallion

Dept. Econ., Georgetown University

- Governments and development agencies throughout the world are trying to make better public spending decisions, esp., more results based.
- They are turning for help to evaluation data and methods.

Are current approaches up to the task?

If not, how can we do better?

Two main messages and
10 recommendations for practice

Message 1: *Policy makers remain poorly informed about what works and what does not in part because:*

- *we do too little evaluation;*
- *we often evaluate the wrong things; and*
- *we often do it the wrong way.*

Knowledge market failures

- Imperfect information about the quality of the evaluation.
  - – Development practitioners cannot easily assess the quality and benefits of an evaluation, to weigh against the costs.
  - – Short-cut non-rigorous methods promise quick results at low cost, though rarely are users well informed of the risks.
- Externalities: Benefits spillover to future projects/policies.
  - – Current individual projects hold the purse strings.
  - – Project manager will not take account of the external benefits when deciding how much to spend on evaluation.
  - – Larger externalities for some types of evaluation (first of its kind; “clones” expected; more innovative)

=> *Without central mandate and support, we will under-invest in evaluation*

Biases in what gets evaluated

- We evaluate a non-random sample of projects/policies.
  - Selected by fashions/favorite methods/TTL prefs.
- We evaluate assigned programs: participants+nonparticipants
  - Programs with large spillover effects and sectoral/economy-wide programs get less attention

+ We evaluate short-lived programs

  - Far easier to evaluate an intervention that yields its likely impact within one year (say) than one that takes many years.
  - Credible evaluations of the longer-term impacts of (for example) infrastructure projects are rare.
  - We know very little about the long-term impacts of development projects that do deliver short-term gains.

=> Knowledge is skewed toward projects with well-defined beneficiaries and quick results

Interaction effects are often ignored

- The (possibly big) bias we don’t much talk about.
- Assessing a development portfolio by evaluating its components one-by-one and adding up the results assumes that there are negligible interaction effects.
  - Yet the success of an education or health project (say) may depend crucially on whether infrastructure or public sector reform projects (say) within the same portfolio have also worked.
  - Indeed, the bundling of (often multi-sectoral) components in one portfolio is often justified by claimed interaction effects.
- If the components interact positively (more of one yields higher impact of the other) then we will overestimate the portfolio’s impact due to double-counting; negative interactions yield the opposite bias.

Biases in how we evaluate

- Obsession 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.
- Social experiments (randomization) can be an important element in the menu of methodological tools.
- However, randomization is only feasible for a non-random sub-set of policies and settings.
- Exclusive reliance on social experiments will make it even harder to address pressing knowledge gaps

Better idea: randomize what gets evaluated and then chose a method appropriate to each sampled intervention, with randomization as one option when feasible.

Message 2: *We can do better, but there are some real challenges in both analytics and administration/implementation.*

10 recommendations for practitioners and policy makers:

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

- Policy relevant evaluation must start with interesting and important questions.

- This may seem obvious, but the reality is that many evaluators start with a preferred method and look for questions that can be addressed with that method.

- Opportunistic RCTs are now common.

- By constraining evaluative research to situations in which one favorite method is feasible, our efforts exclude many of the most important and pressing development questions.

Standard methods don’t address all the policy-relevant questions

- *What is the relevant counterfactual?*
  - “Do nothing”: that is rare; but how to identify relevant CF?
  - Example from workfare programs in India (do nothing CF vs. alternative policy)

- *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 <u>net</u> 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

- Ethically benign evaluations do not change how the program works in practice.
  - If the program is deemed to be ethically acceptable then this can be presumed to also hold for the method of evaluation.
- Ethically contestable evaluations alter the program's (known or likely) assignment mechanism—who gets the program and who does not—for the purpose of the evaluation.
  - Then the ethical acceptability of the intervention does not imply that the evaluation is ethically acceptable.

RCTs are ethically contestable by design

- Scaled-up programs almost never use randomized assignment.
- Pilots (using NGOs) can often get away with methods not acceptable to governments accountable to voters.
- The RCT for a potential public program has a different assignment mechanism to the program, and this may be contested ethically even when the full program is ethically acceptable.

Asymmetric information again

- Key ethical problem: Deliberately denying a program to those who need it and providing the program to some who do not.
- 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.

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

However, ethics is a poor excuse for lack of evaluative effort

- Good ends can sometimes justify bad means. It is ethically defensible to judge processes in part by their outcomes; indeed, there is a long tradition of doing so in moral philosophy, with utilitarianism as the leading example.
- It is not inherently “unethical” to do a RCT that knowingly withholds a treatment from some people in genuine need, and gives it to some people who are not, as long as this is deemed to be justified by the expected welfare benefits from new knowledge.

Far more problematic is either:

1. Any presumption that an RCT is the <u>only</u> way we can reliably learn for the purpose of making better policies.
   - That is plainly not the case, as anyone familiar with the full range of (quantitative and qualitative) tools available for evaluation knows.
2. Any evaluation for which the expected gains from new knowledge cannot reasonably justify an ethically-contestable methodology.
   - This is a judgment call. But it must be addressed explicitly and openly.
   - Only if we are sufficiently ignorant about the likely gains relative to costs should we evaluate further.

3. Take 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.
  - Weak methods of controlling for observable heterogeneity including *ad hoc* (linear, parametric) models of outcomes.
  - Too little attention to selection bias based on observables.
  - Arbitrary preferences for one conditional independence assumption (exclusion restrictions) over another (conditional exogeneity of placement)

We 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 if the program works the way it is intended to work.

Example 2: Southwest China Poverty Reduction Program

- displacement of local government spending in treatment villages => benefits go to the control villages
- substantial underestimation of impact
- Model implies that true DD=1.5 *x* empirical DD
- Key conclusions on long-run impact robust in this case

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.
  - We may well be evaluating the wrong project!

- Flypaper effects
  - Impacts may well be found largely within the "sector".
  - Example for Vietnam roads project: fungibility within transport sector, but flypaper effect on sector.
  - Need for a broad sectoral approach

6. Fully explore impact heterogeneity

- Impacts will vary with participant characteristics (including those not observed by evaluator) and context.

- Participant heterogeneity
  - Interaction effects
  - Also essential heterogeneity, with 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) =>

Impact heterogeneity cont.,

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
  - Systematic covariates (e.g., inequality within villages)

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 change the intervention
- *Outcomes change:*
  - 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 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 and Smith)

- *The RCT has evaluated a different program to the one that actually gets implemented nationally!*

Example of randomization bias

- Two types of people (1/2 of each):
  - Type H: High impact; large gains (G) from program
  - Type L: Low impact: no gain
- Evaluator cannot tell which is which
- But the people themselves can tell (or have a good clue)
- Randomized pilot with small incentive to participate:
  - Half goes to each type with full compliance (epsilon incentive)
  - Mean impact=G/2
- Scaled up program without the incentive:
  - Type H select into program; Type L do not
  - Mean impact=G
- The RCT leads us to substantially underestimate the benefits from this program!

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

In understanding impact, Step 9 is key =>

9. Don’t reject theory and structural models

- Standard evaluations are often “black boxes”: they give policy effects in specific settings but <u>not</u> 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.

Examples

- *PROGRESA* (Attanasio et al.; Todd & Wolpin)
  - Modeling schooling choices using randomized assignment for identification
  - Budget-neutral switch from primary to secondary subsidy would increase impact
- Agrarian reform in Vietnam (Ravallion and van de Walle)
  - Structural model of economy with and without reforms.
    - Living standards model + land-allocation model
  - Calibrated using econometric models of key behavioral and political economy relationships
  - Grounded in historical and qualitative understanding of context.

10. Develop an evaluation culture with strong institutional capabilities

- Strive for a culture of evidence-based evaluation practice.
  - China example: “Seeking truth from fact” + role of research
- Reinforced by strong budget laws.
- An independent central unit in government should decide what gets evaluated and provide quality assurance
  - Example of *Coneval* in Mexico: a model being observed and/or adapted by many countries in LAC and the world.
- Evaluation is often a natural add-on to the government’s <u>sample survey unit</u>. But private data and evaluation capabilities (esp. universities) will be needed too.
- <u>Open access to data and M&E results is crucial</u>. Laws on Access to Information have helped.

Coordination and integration are problems in all countries

- Administration matters! Poor planning, lack of coordination, weak information flows and miss-aligned incentives can undermine the potential for results-based budgeting.
- The wrong things get evaluated (lack of strategy) and the feedbacks to budget decisions are often too weak.
- Portfolios are rarely evaluated; missing integration across projects. Administrative/ministerial silos inhibit effective strategic evaluation efforts.
- Weak integration and poor coordination is too common, both vertically and horizontally.

In summary

There are significant gaps between what we know and what we want to know about the impacts of public programs.

These gaps stem from distortions in the market for knowledge.

Standard approaches to evaluation are not geared to addressing these distortions and consequent knowledge gaps.

But we can do better!

Thank you for your attention!