International Workshop on Impact Evaluation for Development
DAC-OECD, Paris

November 15, 2006

Some Issues and Questions Identified in JBIC's Impact Evaluation Studies

Seiro Ito
JBIC/IDE-JETRO

Contents

1 JBIC Experiences on Impact Evaluation

1. India: role of infrastructure in poverty reduction.
   - Outcome measure: income. Headcount ratio and poverty gap ratio in West Bengal.
   - Intervention: large-scale infrastructure projects (4 power plants, 1 port).
   - Method: poverty analysis macroeconomic simulator (PAMS), a CGE model with a parameterized income distribution function. Simulated possible policy impacts using *ex ante* information; 'with' case is counterfactual and 'without' case is factual.
   - Data: pre-intervention information only, used 55th round of National Sample Survey.

2. Peru: improvement of living environment and livelihood in poor communities.
   - Outcome measure: under-five mortality rate, anthropometrics, individual time use, household income, school enrollment rates in 8 departments of Peru.
   - Intervention: support for Social Fund. Small-scale infrastructure (water supply, rural road/bridges, community electrification).
   - Method: theory-based and propensity score matching.
   - Data: baseline survey of 300 HHs (without household IDs, so a panel could not be constructed), post-intervention survey of 2400 HHs, community questionnaire.

3. Bangladesh: impact of Jamuna Multipurpose Bridge Project on poverty reduction.
   - Outcome measure: income. Headcount ratio and poverty gap ratio in 5 communities of nearby project site.
   - Intervention: large-scale infrastructure project (Jamuna Multipurpose Bridge).
   - Method: theory-based, 'with/without' analysis, and some regressions.
   - Data: baseline used the existing survey data collected by Prof. Bayes (1146 HHs). Post-intervention data was collected by BIDS/IRRI (1585 HHs).

TABLE 1: SUMMARY OF IMPACT EVALUATION STUDIES AT JBIC

Country	India	Peru	Bangladesh
Outcome	income	education, income, health	expenditure, etc.
Intervention	large-scale infrastructure projects	small-scale infrastructure projects (Social Fund)	large-scale infrastructure projects
Methods	simulation (PAMS)	propensity score matching, theory-based	OLS, theory-based
Total Cost	USD 202,000	USD 303,736	USD 21,000
Time	Feb.05 - Oct. 06	Mar. 06 - Jan. 07	Jun.06 - Dec 06
Data	census	survey of 2400 HHs, community questionnaire	secondary data
Total M/M	15M/M	18M/M	2.7M/M
Evaluator	TERI-ICF (India)	GRADE (Peru)	Prof. Bayes, Jahangirnagar U. (Bangladesh)

2 Issues Identified and Open Questions

1. Localization of impact evaluation is limited due to insufficient evaluation capacity development (ECD).

   ✎ Top JBIC financing has been on infrastructure.

   ✎ JBIC considers evaluation should ultimately be done by the partner goverment.

   ✎ JBIC's research department activities has been limited in impact evaluation, and JBIC has been outsourcing to consultants and academia.

   ✎ JBIC mostly completes studies after a few years of project implementation.

   → JBIC's compromise over quality, costs, and time:

   - Use both local (data collection and processing, interpretation of local data) and international experts (technical assistance, advisory, quality control).

   - Select a few programs for more rigorous statistical evaluation, leave the others for descriptive (or theory-based) evaluation.

   - Limited evaluation duration for infrastructure projects.

   - Most studies lack the proper baseline because evaluation is being planned independently at the Evaluation Department.

   ⇨ Who should be the evaluator? International or local? Internal independent evaluation department or academia?

   ⇨ More formally, what are the evaluation goals? How should we divide up the evaluation tasks between these agents, under the budget and time constraints, to best achieve these goals? Can we derive any rules?

   - goals:
     - ① being accountable to the taxpayers
     - ② learning and implementing more cost-effective policies
       - i. creation of knowledge (need: capacity, local context)
         - – estimating consistent treatment effects (capacity, local context)
         - – identifying causal chains (local context)
       - ii. dissemination and use of knowledge (need: donor resources)

   - capacity and costs:
     - – local (less) vs. international (more)

   - knowledge of local context:
     - – local (more) vs. international (less)

   - knowledge creation and sharing incentives:
     - – internal (organization-specific) vs. research/academia (general)

- accessibility to donor resources/decision making:
  - internal (easy) vs. research/academia (difficult)

2. Data collection is often compromised in the routine time frame of project cycle. Budget and time constraints define the scope of (or how much can be done with) evaluation.

   ⇨ What percentage of project costs should be allocated to evaluation?

     - In principle, costs should ultimately be shared with the partner goverment, but that is not practical.

     ① Private optimum: NPV of all marginal increase in treatment effect at JBIC, given other doner's evaluation studies=NPV of all marginal costs of program evaluation at JBIC.

     ② Social optimum: NPV of all marginal increase in treatment effect on the earth=NPV of all marginal costs of program evaluation on the earth.

       - Proportions can be different among the donors given the different cost structure.

       - In principle, each donor should follow its comparative advantage in evaluation.

       - But donors must consult more closely with each other to exploit the complementarity in knowledge creation.

   ⇨ How do we prepare the baseline in advance under the project implementation cycle? There are costs and benefits of:

     ① Universal baseline: all projects collect the simple baseline of temporal information.
       - Costly, prolongs project cycle, less rigorous.
       - No selectivity in evaluation. Supposedly done in the appraisal phase, so no extra costs, in theory.

     ② Selective baseline: stratified random selection of projects for extensive data collection.
       - Effort bias due to announcement prior to implementation, selectivity.
       - More rigorous evaluation.

   ⇨ Who should provide the baseline?

     - Ideally, partner government in conjunction with project appraisal, but they are capacity and resource constrained.

     - Donor practically works with the partner government as a part of evaluation capacity development.

   ⇨ How can insufficient baseline data be coped with *ex post*?

     ① Use of secondary information.

     ② Use of pipeline matching, provided that implementation order is uncorre-

3

lated with unobserved community heterogeneity (unlikely, if not deliberately randomized).

      ③ Poverty map type of smoothing technique for estimation of outcome measures, provided regressors are available for the study area.

      → They all need more variables to control for heterogeneity, thus costlier than panel data.

3. Even with nonbinding time and budget constraints, type and scale of the public intervention (e.g., violation of SUTVA) makes the standard evaluation design inappropriate.

     ⇨ How can treatment/control groups be constructed in large-scale infrastructure projects? Can we reasonably handle the GE effects by modifying SUTVA?

        • Controls are even harder to construct.

        • Simulation can ask interesting questions but imposes strong assumptions on the structure of model.

        • One can estimate the treatment intensity by imposing flexible, spatially smoothed structure. Such intensity can be endogenous to placement, so one needs a fairly accurate placement rule.

4. Even without any constraints and under SUTVA, it is often not straightforward to draw policy implications from the estimated results. Theory-based method presents some plausibility but cannot establish the causal relationships.

     ⇨ What do we want to know to begin with?
       ① Size and distribution of impacts (total derivatives): estimation.
       ② Bottlenecks and catalysis (partial derivatives): finer estimation strategy with finer and further experiments/instruments, or resort to theory-based method. One can mix the two, but there is no formal way to do so.

     ⇨ Which programs should be chosen for statistical evaluation?

        • But project implementation is purposefully chosen, so the universe is a selected sample.

     ⇨ Backing up one step more, which programs should be chosen for implementation?
       ① Local political economy dictates to considerable extent.
       ② A 'causal' poverty map: poverty measures regressed on instrumented policy and other exogenous variates. This gives a guidance on priority to policy makers, and also provides a placement rule for researchers.

5. One needs to use scientific/objective measurement for the health- and hygiene-related variables due to errors in self-reporting (Strauss and Thomas, JEL 1998).

     ⇨ What kind of community/household data should be collected through scientific measurement?

① Self-reported measurements that may be correlated with unobservable individual/community heterogeneity.

② Measures that are difficult to perceive or conceptualize.

⇨ How do we collect them?

- In collaboration with local NGOs and partner health ministry.

- In collaboration with international agencies for standardization of measurement practice.

6. When a census is not available, sample is sometimes stratified on unknown ground such as distance from the village center.

⇨ How can the sampling strategy be set (stratification, number of samples) before the survey?

- Stratification depends on between what groups we want to see the differential impacts. Rich vs. poor; landed vs. landless; literate vs. illiterate; cohorts; ethnicity; religion.

- Problem is that the distribution of strata is not always available *ex ante*.

- Relevant choice of grouping over whom we contrast the differential impacts is not always known to international experts. Local context is valuable in choosing the stratification.

- Number of samples can be decided with power calculations, which requires small test samples or other secondary data.

7. There are almost always other selective interventions.

⇨ How should we measure the impacts of a set of multiple interventions and policies (multiple treatments)?

⇨ How can we separate out (orthogonalize) the impacts of other interventions?

① Randomization.

② IVs that are uncorrelated with other interventions and uncorrelated with individual heterogeneity. (Is there any?)

3 Summary

1. Prior to evaluation, one needs to consider:
   - evaluation goals
   - time and budget constraints
   - available human resources and information
   - coordination with other donors
   - knowledge dissemination and feedback strategy

2. Baseline data can be collected as a routine in two ways:
   - universal: costly, tedious, less rigorous, yet no selection.
   - selective: less tedious, more rigorous, yet selectivity and effort bias.

3. Construction of controls is limited and costlier with cross-section data. So better getting the baseline.

4. Large-scale projects are more difficult to evaluate, requiring simulation based on strong assumptions or estimation of endogenous treatment intensity.

5. Estimation of causal poverty maps can serve as a guidance to policy makers and a placement rule to researchers.

6. Health- and hygiene-related variables need to be measured objectively, in collaboration with local NGOs, partner government, and international agencies.

7. Knowledge of local context is relevant in the choice of stratification.

8. Bottlenecks and catalysis, or causal chains, and multiple treatments can be identified with estimation with further and finer experiments/IVs. But the scope is limited.