

# **Learning to Evaluate the Impact of Aid**

**Seiro Ito, Nobuyuki Kobayashi and Yoshio Wada**

**NONIE WORKING PAPER NO. 6**

**January 2008**

## What is NONIE?

---

Nonie is a network of networks for impact evaluation comprised of the DAC Evaluation Network, The United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and a fourth network drawn from the development evaluation associations (AfrEA, IOCE, IDEAS, ReLAC, and IPEN). Its purpose is to foster a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation.

To this end a task team has been constituted and tasked with the following activities:

1. Preparation of impact evaluation guidelines
2. Agreeing collaborative arrangements for undertaking impact evaluation, leading to initiation of the program
3. Developing a platform of resources to support impact evaluation by member organizations

## **NONIE Working Papers**

NONIE working papers present conceptual papers and impact evaluation findings. They may have been published elsewhere, e.g. as government or agency reports, but are included in the NONIE series to increase dissemination. Feedback on papers via the NONIE website is welcome.

# Learning to Evaluate the Impact of Aid

Seiro Ito, Nobuyuki Kobayashi and Yoshio Wada

## 1 Introduction

Recent cross-country panel regression analyses have revived the debate over the effectiveness of development aid on economic growth (World Bank 1998; Burnside and Dollar 2000; Easterly *et al.* 2004; Roodman 2007). These arguments have been fuelled by results-based management of development aid, such as the Paris Declaration, and naturally lead to a call for more rigorous impact evaluation at the project level. The importance of impact evaluation is frequently stressed by the international aid community (World Bank 2006 and Savedoff *et al.* 2006; Asian Development Bank 2006; Banerjee 2007).

Bilateral development aid institutions, however, have been slow to respond to requests for impact evaluation despite being in general agreement with the targets of the Millennium Development Goals and their stated commitment to pay due attention to aid effectiveness. Why are bilateral development aid institutions slow to accept the importance of impact evaluations? In order to answer this question, this article examines the experience of the Japan Bank for International Cooperation's (JBIC) rigorous impact evaluation of Japan's ODA projects and draws some lessons relevant to other donor institutions. We show that the particular features of Japan's aid, and the aid environment in general, are impediments to adopting rigorous impact evaluations on a full scale. We also examine what the introduction of rigorous impact evaluation would mean for bilateral aid.

While the proponents of rigorous impact evaluations are drawing growing attention in the evaluation community, one rarely sees a discussion of how one can learn from past aid programmes in a systematic manner. We point out that rigorous impact evaluation alone cannot draw lessons from past aid programmes, and argue that we must understand the mechanism that produced the measured impact to better predict future aid impacts. Summative aspects of rigorous impact evaluations have been discussed intensively while formative aspects have long been neglected. Based on the Bayesian statistics/econometrics literature, we will discuss one method of inferring such mechanisms with data. We show the need for sharing data and evaluation experiences among evaluators, and the need for the evaluation community to work more closely with the research community.

After the introduction to this article, Section 2 examines the characteristics of Japan's ODA and describes how the aid community in Japan reacts to requests for rigorous impact evaluation. Section 3 introduces evaluation designs, evaluation results, and other findings of rigorous impact evaluation recently undertaken by JBIC in Bangladesh and Peru. In Section 4, we use the Bayesian framework to show how rigorous impact evaluations can feed back into the aid projects, and make the case for closer collaboration between the international aid community and the research community. Section 4 concludes.

## **2 Japan's development aid and rigorous impact evaluation**

### *2.1 Japan's ODA characteristics and their implications on impact evaluations*

The evaluation of Japan's ODA takes place at three levels – the policy, programme, and project levels – and is conducted by three institutions – the Ministry of Foreign Affairs, JBIC, and JICA (Japan International Cooperation Agency). According to the ODA Evaluation Guidelines (Ministry of Foreign Affairs (2003)), the objectives of evaluation are 'Feedback to ODA management' and 'Accountability.' The impact of development aid is perceived to be one of the criteria for evaluation as mentioned in the 'JBIC Ex-post Evaluation Report on ODA Loan Projects 2006'. Methodologically, however, the standard evaluation assesses overall project performance by process evaluation only and rigorous impact evaluations are rarely performed or officially incorporated in the Plan-Do-Check-Action (PDCA) cycle of Japan's ODA operations.

The DAC peer review reports (2003, 2007) identify five characteristics of Japan's ODA that affect evaluation. First, Japan's ODA clearly focuses on Asian countries for historical and geopolitical reasons. This is shown by a quote from Japan's Official Development Assistance Charter, 'Priority Regions: Asia, a region with close relationship to Japan and which can have a major impact on Japan's stability and prosperity and is a priority region for Japan.' A sizeable portion of Japan's ODA is allocated to China, Indonesia, the Philippines, Thailand, and India. These countries are very populous, ranging from one hundred million to more than one billion, and are growing at faster rates than other economies.

Secondly, Japan's ODA supports provision of large-scale economic and social infrastructure, such as roads/bridges, power stations, transmission lines, water supply and so on. High population density, high economic growth, and accompanying rapid urbanisation in Asian countries have created substantial demand for these projects. As noted in Howard White's article (in Banerjee 2007), randomisation is impossible for this type of aid.

Thirdly, large-scale ODA projects tend to be financed by loans while the small-scale projects are financed by grants. Creating the counterfactual is difficult for Japan's development assistance because of its great emphasis on large scale infrastructural projects. The impact evaluation of the Jamuna Bridge in Bangladesh described in the next section illustrates these challenges.

Fourthly, in Japan's ODA, each of the three key players conducts separate and independent evaluations for each aid modality: grants, loans, and technical cooperation. There is no practice of conducting joint evaluation of all modalities of ODA (grant, loan, and technical cooperation) within a certain sector or country, and rigorous impact evaluation on a group of small projects is rarely conducted.

Fifth, the Japanese government maintains a clear principle of request-based commitments of aid. Unless asked by the partner government, they will not provide ODA. This approach has the advantages of accommodating partner government preferences and strengthening the project or programme ownership, and limiting the aid agency to negotiations only with the partner's central government. However, this also has an adverse implication for impact evaluations, as most of the projects are chosen by the policymakers leading to the likelihood of placement bias in a naïve impact estimation. We turn to this problem in the next section.

## *2.2 Implications of global trends*

Since the late 90s, there has been a steady shift in the types of projects Japan funds. The increased support for small-scale rural infrastructure, such as rural road and community waterworks projects, and small-scale participatory projects, such as microfinance and social funds, reflects the increasingly decentralised decision-making on development in the East and South Asian countries. In many cases, these small-scale projects are dispersed across rural areas, and are selected and implemented by the local governments in those areas, based on applications from local communities. This creates a self-selection process in the project implementation, which must be accounted for in the rigorous impact evaluation studies. Microfinance, which is rapidly expanding throughout Asia, is also subject to selection bias. Econometric techniques for controlling such bias have been developed but only recently applied to aid-supported interventions. These changes call for technical expertise in controlling biases, and naïve impact evaluations that ignore them are considered inaccurate.

So far, it is fair to say that rigorous impact evaluation has not played a vital role in Japan's ODA. However, JBIC is beginning to institutionalise rigorous impact evaluations and plans to conduct a greater number of impact evaluation studies. The impact evaluations on FONCODES, a Peruvian social investment fund, and another on the Jamuna Bridge, described in the next section, reflect JBIC's response to the growing global demand for rigorous impact evaluation.

## *2.3 Steps for rigorous and frequent impact evaluations*

JBIC requires its evaluators to conduct a beneficiary survey in all *ex post* evaluations of ODA loan projects. In the beneficiary survey, however, the construction of a counterfactual is not required. In addition, there are few baseline surveys to collect information of the treated group at the household level, and even fewer baseline surveys cover control groups. The resulting evaluations are likely to be plagued with various kinds of biases, so one must impose implausibly strong assumptions and assume away the biases. This is clearly unsatisfactory. If the intervention is randomised, difference-in-differences (double differencing) allows estimation of impact under plausible assumptions. Randomisation, where possible, is the key step for introduction of rigorous impact evaluations to get the less bias-ridden impact estimates.

Among the difficulties attached to incorporating baseline data collection, we note, is the vertical segmentation of project flows in Japan's ODA. In Japan's ODA projects, feasibility studies are conducted by the technical cooperation agency, JICA, not by JBIC which finances ODA projects. This bureaucratic separation of technical assistance from financial assistance is a significant barrier to the introduction of rigorous impact evaluation in Japan.<sup>1</sup>

## **3 JBIC's experiences in impact evaluation**

### *3.1 The impact evaluations in Bangladesh and Peru*

JBIC has conducted several impact evaluations on ODA-funded projects. These evaluations employed quasi-experimental designs. This section introduces the impact

evaluation on the Jamuna Bridge as a large-scale infrastructure example and on FONCODES as a dispersed area, small-scale infrastructure example.

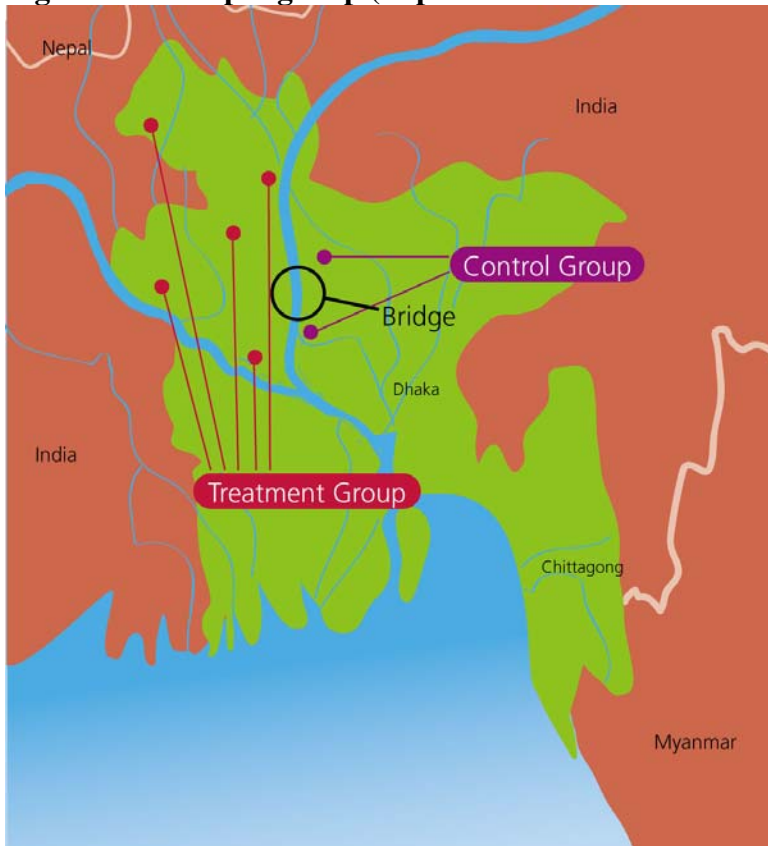
### *3.1.1 Bangladesh: the impact evaluation on the Jamuna Bridge<sup>ii</sup>*

Jamuna Bridge is one of the largest infrastructure projects in South Asia, costing approximately US\$960 million. JBIC, ADB (Asian Development Bank), and the World Bank jointly financed this project, which constructed the 4.8km bridge and 30km of approach roads. The Jamuna River obstructs smooth traffic between central Bangladesh and the northwestern part of the country. Before the completion of the Jamuna Bridge, the major mode of transportation was ferry boats which took at least two hours to cross the river. Due to the limited capacity of the boats, there was a prolonged waiting time for river crossing, in particular for freight trucks. The Jamuna Bridge presumably has had a substantial effect on commercial and economic movements within the country, in particular in the movement of agricultural products from the northwest to the central region and movement of industrial goods from the central to the northwest region.

The impact evaluation on the Jamuna Bridge uses household income, agricultural production, and agricultural inputs (such as pesticides and fertilisers) as impact indicators. As baseline data at pre-intervention were available, the evaluation uses the difference-in-differences (DID) estimator. It is a reasonable assumption that the treatment groups had no opportunities to participate in the decision regarding the placement of the Jamuna Bridge. For this reason, the evaluation does not have to consider selection bias.

Five communities on the northwestern side of the Jamuna River are defined as the treatment group while two communities on the eastern side of the river are the control group. From the universe of survey data from the Bangladesh Institute of Development Studies and that of another study by the International Rice Research Institute, the treatment and control groups were selected randomly from the village-level population lists. In the 1997/98 season, before the bridge began to operate, a census of village households collected data on 1,585 households. In 2003/04, the same households were visited with the same questionnaires. The number of households surveyed, however, decreased to 1,146 in 2003/04 due to migration and river erosion. The number of households decreased further when some of the households had to be omitted due to low quality of the respondents' answers.

**Figure 3.1: Sampling map (impact evaluation of the Jamuna Bridge)**



DID is now considered one of the most reliable estimators to measure policy impacts in the absence of explicit randomisation. Our experience, however, shows the difficulty in impact evaluation even when researchers are equipped with the baseline data. The high attrition rate, 28 per cent over six years or an average of 5.4 per cent per annum, poses a potential problem in the panel data analysis. If the migration is related to the unobservables that may affect the outcome of interest, then we have a sample selection problem: the households that fall out of the data set are those that earn less income and are less well-off, leaving the higher-ability households in. Hence the impact estimate is upwardly biased.<sup>iii</sup> Given we have only cross-section data for the attributed households, one cannot test if the attrition is due only to observables. However, the limited size of the study budget did not allow the evaluator to track down the migrated households to determine the reasons for migration. Thus we had to assume that the sample selection problem is nonexistent, in other words, they are 'missing at random (MAR)' after conditioning on the observables.

The impact evaluation of the Jamuna Bridge brings to light several findings and policy implications. The treatment groups grew more high-value crops. The Jamuna Bridge presumably decreased distribution costs for these high-value but more perishable crops and stimulated farmers to grow them. Storage facilities for these crops would help farmers to earn a better return. Furthermore, one of the unexpected findings is a change in the credit sources. The treatment groups relied more on NGO and less on traditional credit sources (i.e. moneylenders and friends/relatives). Although a similar pattern was observed in the control groups, the change was more significant among the treatment

groups. Cheaper costs for logistics improved not only access to product markets but also to credit markets.

### *3.1.2 Peru: impact evaluation of the social investment fund<sup>iv</sup>*

FONCODES, a Peruvian social investment fund, supports small-scale infrastructure projects in rural areas. JBIC provided ODA loans to FONCODES in order to finance a wide variety of subprojects in eight provinces. The amount of investment of a typical FONCODES subproject ranged mostly from US\$30,000 to US\$50,000 and the ODA loans that JBIC provided enabled FONCODES to support more than 1,000 communities. The application process of FONCODES subprojects is community-driven. Community members organise a community meeting, select a type of subproject, and jointly submit an application form to FONCODES. FONCODES reviews these application forms and selects subprojects to be financed in consideration of various factors such as poverty level, size of investment, and FONCODES' past support to the community.

Three types of subproject (rural road/bridge, water supply, and rural electrification) were chosen for the FONCODES impact evaluation from several lines of support. As these are subprojects, we can draw upon a sufficient number of communities. Impact indicators were set in consideration of their relevance to the Millennium Development Goals. The participatory nature of the selection process for the subprojects requires removal of selection bias in the construction of a counterfactual. Baseline data were available in only a few places at the community level. This led the evaluators to use propensity score matching at both community and household levels as the way to control the self-selection bias. The variables used in propensity score matching are geographical variables, community characteristics, and poverty assessment by the government.<sup>v</sup>

**Figure 3.2: Sampling map (the impact evaluation of the social investment fund)**



A household survey targeted information collection from 2,240 households in 224 communities (10 households per community) in seven departments, including both the treatment and control groups. A subproject selection model of the communities was estimated for each subproject with probit, using the pre-intervention variables of pre-Census 1999 data and the Poverty Map of 1996 as regressors. For each line in each department, non-intervened communities were identified as the potential controls for all the intervened communities. A community with the closest value to the fixed propensity score radius was chosen as the control community. This exercise was carried out for all the communities and for the three types of subproject in each department.

From the impact evaluation on FONCODES, the evaluation findings on water supply subprojects suggest that household characteristics play a crucial role in the impact on health conditions. At the community level, the main positive results found are reduced daily time required to collect water, more expenditure on hygiene products like soap, and improved perception of water quality. No significant impact was found for indicators on health conditions like incidence of diarrhoea in children 0–6 years old, or nutrition status among children. At the household level, however, two health indicators – incidence of diarrhoea and that of skin diseases among children 0–6 years old – show positive results. The difficulty in detecting the impact on health conditions by community-level matching

implies that household characteristics play a crucial role in family health and, therefore, knowing what happens inside the household is an essential step in implementing water supply projects as well as in impact evaluation.

### *3.2 Factors creating bias in the estimation of impacts*

#### *3.2.1 Large-scale infrastructure projects*

It is a conventional assumption that the closer the treatment and control groups reside to each other, the more similar their group characteristics will be, including the unobservable community characteristics. So whenever feasible, the evaluators try to find control groups in proximity to the treated group. The effects of large-scale infrastructure, however, spill over from these areas. This makes it inappropriate to select both the treatment and control groups from proximate areas, and forces the evaluators to find a more remote control group. We therefore will be caught between a rock and a hard place of the similarity concern versus the spillover concern. The evaluators may not be able to find the counterfactuals in the data, regardless of the geographical coverage of the survey. In the impact evaluation of the Jamuna Bridge, we assumed the effect of the project on the control group on the eastern side of the river to be negligible. But this disregards the fact that the Jamuna River affects the distribution of goods throughout the entire country.

Technically, large-scale projects have an equilibrium effect that poses a challenge to rigorous impact estimation. Large projects affect prices, so the counterfactuals cannot be observed directly from the data because people's choices change along with the price changes. Hence the counterfactual must be computed using prices in the absence of the project. For example, in the Jamuna Bridge case, the construction of the bridge may have tightened the ties between the rural and urban labour markets. This increased the demand for labour in the rural areas, and raised the income of labourer households. So even in the control villages that seemed not to have been directly affected by the bridge, income increased. This may give a downward bias to the simple impact estimate, as we cannot observe in the data the rural income in the absence of greater integration of the rural labour market with its urban counterpart. The construction of the counterfactual under the general equilibrium effects is not easy, and requires simulation exercises. This means that the evaluator not only needs to work with the researchers, but also needs to impose more assumptions on the economic relationships to keep the computation feasible.

#### *3.2.2 Dispersed small-scale infrastructure projects*

On the other hand, the impact evaluation of dispersed, small-scale infrastructure projects faces different types of challenges for accurate estimation of impact. Dispersed small-scale infrastructure projects usually employ a community-driven application from potential beneficiaries. This results in self-selection and heterogeneity of interventions, both of which hinder the accurate estimation of impacts. The first issue, the self-selection bias, can partly be dealt with using the propensity score matching method. However, mimicking the application process in the estimation of the propensity score is often more difficult than it looks because the informal placement rules and the internal dynamics within the community during the application process are usually unobservable. One can sometimes use local expertise to better deal with this issue. For example, proximity to

FONCODES field offices was included in the estimation of FONCODE subproject selection probabilities after the evaluators learned that the FONCODES promotion activities rarely reach faraway communities. A second issue, heterogeneity of intervention, is the direct result of decentralised decision-making because the design of subprojects will reflect local demands. Each type of subproject consisted of several subcategories with minor differences. The rural road/bridge subprojects include sidewalks within a community, unpaved roads or bridges that reach other communities, and even small bridges at the piers of small river ports. Some roads are only for pedestrians and some are for motor vehicles. It is possible to cope with heterogeneity of intervention by further division of subprojects, but only at the cost of having a smaller sample size.

### *3.3.3 Request-based commitments*

Because the Japanese government provides ODA on a request basis, impact estimation is also constrained by the placement bias. It is reasonable to think that the policymaker places the project where it best suits his objective. If the placement is correlated with unobservable factors such as the ability of people to earn income that may influence the outcome of interest, the availability of services provided by the project will be correlated with the unobservable, which we shall call,  $a$ . Such a presumption is plausible when the policymaker has the information that is relevant to placement but not available to the evaluators. The policymaker may want to choose areas reflecting higher  $a$  values, so the impact on income will be greater. The treatment thus may not be uniform across the population, but is given more to the people with higher  $a$ s. We thus have an endogeneity problem which results in an upwardly biased impact estimate because the treatment is given primarily to the people with above average ability. In the FONCODES case, unfortunately, we could not obtain information on  $a$  that might have influenced the policymaker's decision. We had no other choice but to assume away the placement bias.

### *3.3 The current use of evaluation findings at JBIC*

Rigorous impact evaluation is a powerful tool to satisfy the objectives of summative evaluation by providing useful information to policymakers, who then decide whether to continue the intervention or to change the design of future interventions. JBIC currently uses the evaluation findings such as those mentioned above mainly to satisfy calls for accountability. This is another effective use of summative evaluation, as the evaluation findings show the results in a clear-cut manner. JBIC conducts *ex post* evaluations on all projects by using the framework of the DAC Five criteria, some of which ('Effectiveness' and 'Impact') touch upon the aid effectiveness of projects. Nevertheless, the analysis at *ex post* evaluation rarely goes beyond 'Before/After' or 'Planned/Actual' comparison of key performance indicators. Although 'Before/After' or 'Planned/Actual' comparison can show rough results of projects, these suffer from various biases explained above. In understanding aid effectiveness, the rigorous impact evaluation gives a clearer picture of Japan's ODA projects, and should serve the accountability purpose better.

Despite our support for the promotion of internally valid impact evaluation and the development of externally valid knowledge, we do not support discontinuing the formative evaluation efforts. To the contrary, we strongly believe that the formative aspects should remain a key tool. This is because the importance of formative evaluation

in deterring corruption can never be overstated. Take the example of two flagship projects for maternal and child health in India. Unfortunately, it was found that these programmes were afflicted with bid rigging, bribery, use of unqualified products, and forged performance reviews<sup>vi</sup> that resulted in the World Bank halting financing for the programmes for a prolonged period. Sometimes, as in the health projects in India, corruption deterrence can offer greater mileage than devising a better technical design. Given the primacy of corruption deterrence in securing various entitlements (which is especially true in the health sector), summative evaluation must go hand in hand with formative evaluation. The evaluator can give, if necessary, a rationale to the partner government wanting to deter corruption and implement the programmes more transparently.

In the next section, we will show a general methodology for assessing the external validity from the internally valid impact evaluations.

#### **4 A workable framework for external validity**

In the context of Savedoff *et al.* (2006), impact evaluation can be seen as a means to achieve two goals: to provide the taxpayers accountability (achieving internal validity), and to produce the knowledge to maximise per dollar effectiveness (achieving external validity). It is, however, well recognised that the impact evaluation by itself can only reveal the results, not the mechanism that created it. In other words, it only deals with the internal validity. In most of the rigorous evaluation studies, external validity is discussed informally in contrast to its internal counterpart. Savedoff *et al.* (2006) emphasises the importance of achieving external validity, but they are silent on how this can be done.

Why, then, implement rigorous evaluations when they only serve one of the goals? How can we deduce the mechanism from the internally valid evaluation studies? One needs a clear, systematic methodology to synthesise the available evidence to produce the new knowledge. This section aims at providing one such account, drawing on the idea of Bayesian statistics/econometrics literature. It shows that the way to produce useful knowledge about the potential impacts, denoted as  $\Delta\tilde{y}$ , and the mechanism, denoted as  $M$ , working behind such impacts.

*Bayes' rule* states for a model  $M$  and data  $Y$ :

$$\Pr[M | Y] = \frac{\Pr[Y | M] \Pr[M]}{\Pr[Y]} \quad (1)$$

That is, the conditional probability that the model  $M$  is at work after observing the data  $Y$  is equal to the right hand side of (1).

$\Pr[M]$  is a *prior distribution* of  $M$ . It is prior in the sense that it is determined before we observe data  $Y$ . One reasonable choice of prior is a posterior distribution from previous estimation or the consensus of the field. When there is no justification or reason for choosing a particular probability density function, Bayesians sometimes resort to the use of *noninformative* or *flat* priors to escape from the arbitrary choice. A flat prior, in essence, determines all the elements as having an equal chance of happening. It is also noteworthy that as the sample size grows, the posterior will be less dependent on the particular choice of a prior.

$\Pr[Y|M]$  is the probability we observe  $Y$  when the model  $M$  is at work, or the *likelihood* of data, given model  $M$ . This can be computed for each different model. The final bit,  $\Pr[Y]$ , is the probability of observing data. It is not important for our discussion to know what it is, so we will skip the explanation of it.<sup>vii</sup>

To see how the likelihood can be computed for each model, let us consider an example. Suppose that a parent allocates nutrition  $y$  among the children in a household. Presumably, it is an increasing (or a non-decreasing) function of wealth  $W$ . A theory of taste-based intrahousehold gender discrimination (mechanism 1, or  $M_1$ ) may indicate that if the child is a girl, denoted with  $F = 1$  and  $F = 0$  for boys, then less nutrition may be given. Another theory, call it out-of-necessity gender discrimination theory ( $M_2$ ), may tell that it is only when the household is poor that parents may decide to discriminate against daughters in favour of sons, so nutrition is an increasing function of the interaction term  $F \bullet W$ . In a regression form, we have:

$$y = \theta_0 + \theta_1 F + \theta_2 F \bullet W + \theta_3 W + u \quad (2)$$

$u$  is an error term which follows some distribution function  $G(u)$ . The pure taste-based gender discrimination theory  $M_1$  would predict  $\theta_1 < 0$ ,  $\theta_2 = 0$ ,  $\theta_3 = 0$ , and the out-of-necessity gender discrimination theory  $M_2$  would predict  $\theta_1 = 0$ ,  $\theta_2 > 0$ , and possibly  $\theta_3 > 0$ . So the different theories give different restrictions on the parameter space. The likelihood of data given model  $M_1$  is the value of the distribution function of the above equation with restrictions  $\theta_1 < 0$ ,  $\theta_2 = 0$ ,  $\theta_3 = 0$  imposed. So we evaluate the value of function  $G(u|M_1) = G(y - \theta_0 - \theta_1 F)$  with  $\theta_1$  restricted to be some negative value.

If the collected data show that  $y$  is not increasing with  $F \bullet W$ , then a proposed mechanism called the out-of-necessity gender discrimination theory  $M_2$  has less chance to hold, thus it has a smaller power in explaining the data at hand, or  $M_2$ 's prediction of data at hand  $Y$  has a small probability, or  $\Pr[Y|M_2]$  is low. Consequently, using the Bayes' rule, we have a high probability for  $M_1$ , or  $\Pr[M_1|Y]$ . We have created the knowledge that the taste-based gender discrimination may be prevalent *in this study area*. However, we have *not* shown if  $M_1$  is relevant to other areas.

Using the posterior probability for a model  $M$ , we can derive the *posterior predictive probability* of the potential impact  $\Delta\tilde{y}$ . To do so, let us assume that data contains two elements, impact  $\Delta y$  and other explanatory variables  $x$ , or  $Y = (x, \Delta y)$ , and assume for the moment that there are only two models,  $M_1$  and  $M_2$ . Then, posterior predictive probability of  $\Delta\tilde{y}$  given  $x$  and  $\Delta y$  is:

$$\Pr[\Delta\tilde{y} | x, \Delta y] = \Pr[\Delta\tilde{y} | M_1, x, \Delta y] \Pr[M_1 | x, \Delta y] + \Pr[\Delta\tilde{y} | M_2, x, \Delta y] \Pr[M_2 | x, \Delta y] \quad (3)$$

$\Pr[M_i | x, \Delta y]$  is obtained from (1) by replacing  $Y$  with  $x, \Delta y$ ,  $\Pr[\Delta\tilde{y} | M_i, x, \Delta y]$  is the likelihood of potential impact  $\Delta\tilde{y}$  when model  $M_i$  is at work and we have additional variables  $x$  to explain  $\Delta y$ . Note that the potential impact only happens under  $M_1$  or  $M_2$ ,

and the probabilities conditional on each model being at work are given by the likelihood  $\Pr[\Delta\tilde{y} | M_1, x, \Delta y]$  and  $\Pr[\Delta\tilde{y} | M_2, x, \Delta y]$ . The total probability is given by multiplying these conditional probabilities with the probabilities of a conditioning event, or the probabilities that each model is at work,  $\Pr[M_1 | x, \Delta y]$  and  $\Pr[M_2 | x, \Delta y]$ . Hence we have the above.

Analogously, when there are  $I$  models  $M_1, \dots, M_I$ , the posterior predictive probability of  $\Delta\tilde{y}$  given  $x$  and  $\Delta y$  is:

$$\begin{aligned} \Pr[\Delta\tilde{y} | x, \Delta y] &= \Pr[\Delta\tilde{y} | M_1, x, \Delta y] \Pr[M_1 | x, \Delta y] + \dots + \Pr[\Delta\tilde{y} | M_I, x, \Delta y] \Pr[M_I | x, \Delta y], \\ &= \sum_{i=1}^I \Pr[\Delta\tilde{y} | M_i, x, \Delta y] \Pr[M_i | x, \Delta y] \end{aligned} \quad (4)$$

Equation (4) shows how the relevance of  $M_i$  to other areas affects the predictive probability of future possible impacts  $\Delta\tilde{y}$ . In terms of the example above, the greater relevance of  $M_1$  is expressed as a high posterior probability of  $M_1$ , or a large  $\Pr[M_1 | x, \Delta y]$ . So  $\Pr[\Delta\tilde{y} | M_1, x]$  for all the possible values of  $\Delta\tilde{y}$  will receive a higher weight  $\Pr[M_1 | x, \Delta y]$ . Then  $M_1$  gets the larger weights in computing  $\Pr[\Delta\tilde{y} | x, \Delta y]$ .

The procedure described is called *Bayesian model averaging* (BMA) which is useful in evaluating numerous, unknown alternative models. There are several approaches to implementing BMA. One can use the *Markov chain Monte Carlo model composition* (MC<sup>3</sup>) which is the Metropolis-Hastings algorithm applied on a collection of models/mechanisms. Another possibility is to use the *Occam's window* approach which sets up criteria for selecting the models/mechanisms using the posterior odds ratio. As with any Bayesian method, one needs to evaluate the multiple integrals implicit in Equation 2. This adds a considerable computational burden, and its implementation is not easy. See Hoeting *et al.* (1999).<sup>viii</sup> This calls for the evaluators to collaborate with the empirical experts on BMA.

We do not, of course, claim that the above is the only methodology available. A counterpart frequentist version exists, and interested readers are advised to consult it.<sup>ixx</sup> Recently, Todd and Wolpin (2006) have shown another way to verify the mechanism behind the impact. They split the sample into the model training (estimation) and the model testing (out-of-sample prediction) parts, and tested the mechanism derived from the model training part with the data in the model testing part. Their methodology is valid because their testing part is known to have an accurate impact estimate  $\Delta y$ .

Our proposed approach is similar in spirit. Suppose that we have from area 1 data  $x_1$ , the impact estimate  $\Delta y_1$ , and the candidate models  $M_1, \dots, M_I$ , and we want to consider the possible programme impact in area 2. Having collected  $x_1$ , we estimate  $\Pr[M_i | x_1, \Delta y_1]$  with  $x_1$  and  $\Delta y_1$  for each  $M_i$  with (1). We can obtain posterior predictive probability  $\Pr[\Delta\tilde{y} | x_1, \Delta y_1]$  using all the possible models in Equation (4). If we collect the variables  $x_2$  in area 2 prior to the intervention, we can replace  $\Pr[\Delta\tilde{y} | M_i, x_1, \Delta y_1]$  with  $\Pr[\Delta\tilde{y} | M_i, x_2, \Delta y_1]$  by substituting  $x_2$  in place of  $x_1$  in the likelihood  $\Pr[\Delta\tilde{y} | M_i, x_1, \Delta y_1]$ . Then we can re-compute the posterior predictive probability as  $\Pr[\Delta\tilde{y} | x_2, \Delta y_1]$ .

In sum, the steps to follow between  $t$ th and  $t + 1$ th evaluation are:

1. Given  $x_t$ ,  $\Delta y_t$ ,  $\{M_i\}$ , for each  $M_i$ , compute  $\Pr[\Delta\tilde{y}_{t+1} | M_i, x_t, \Delta y_t]$ , derive  $\Pr[M_i | x_t, \Delta y_t]$ , and get  $\Pr[\Delta\tilde{y}_{t+1} | x_t, \Delta y_t]$ .
2. Once  $x_{t+1}$  becomes available, replace  $\Pr[\Delta\tilde{y}_{t+1} | M_i, x_t, \Delta y_t]$  with  $\Pr[\Delta\tilde{y}_{t+1} | M_i, x_{t+1}, \Delta y_t]$  and update the posterior predictive probability to get  $\Pr[\Delta\tilde{y}_{t+1} | x_{t+1}, \Delta y_t]$ .
3. When  $\Delta y_{t+1}$  becomes available through rigorous impact evaluation, fully update to get  $\Pr[M_i | x_t, x_{t+1}, \Delta y_t, \Delta y_{t+1}]$  and  $\Pr[\Delta\tilde{y}_{t+2} | x_t, x_{t+1}, \Delta y_t, \Delta y_{t+1}]$ . With an insight from evaluation study or data, add other candidate models  $M_{t+1}, \dots$  if necessary.

This suggests that it is imperative to pool information on  $\Pr[\Delta\tilde{y} | M, x, \Delta y]$ ,  $\Pr[M | x, \Delta y]$ , or previously collected data  $\{x, \Delta y\}$  with a collection of candidate mechanisms  $\{M_i\}$ . A researcher or an evaluator who has a new data set  $\{x, \Delta y\}$  should be able to update the posterior distributions.

It does not, however, necessarily call for the creation of a new entity nor central planning of evaluation studies. Academia is a good working example of how a decentralised community shares useful information to update knowledge; a website may suffice, provided that evaluation research funding is conditional on posting the data on a website. This also means that an evaluator who is interested in achieving external validity must know the theories  $M$  and their implications for the respective likelihood  $\Pr[x, \Delta y | M]$ . Again, it suggests a scope for fruitful collaboration with the research community.

## 5 Conclusion

Drawing on JBIC's experience, we showed that historical and institutional features of Japan's ODA such as concentration in large-scale infrastructure projects in Asian countries, separation between technical and financial assistance administration, and the principle of request-based commitment impede the swift adoption of rigorous impact evaluations. Randomisation is impossible in large-scale infrastructure projects and the construction of counterfactuals directly from actual survey data may not be appropriate. Bureaucratic separation between technical and financial assistance bars the project management from the evaluator's perspective. Request-based commitment allows a greater placement bias by the policymakers.

We also showed that the global trends of decentralisation and participatory decision-making are posing additional difficulty in estimating impacts consistently. Dispersed small-scale infrastructure projects are often based on local request and decentralised decision-making, and the removal of self-selection bias requires proper understanding of application processes. Decentralised decision-making also results in heterogeneity of treatment.

For accountability reasons, JBIC's impact evaluation has traditionally focused on summative aspects. With rising interest in internally valid impact evaluations, JBIC has started to treat the summative aspects of internal validity as a way to show the effectiveness of the ODA projects. The focus on summative aspects, or a rigorous impact

evaluation, is an important change of direction, but this does not mean that we can neglect the utility of formative evaluation in securing transparency.

Finally, we have pointed out that rigorous impact evaluation alone is not sufficient for learning from past programmes, and argued that we must understand the mechanism that produced the measured impact to better predict the future aid impact. The notion of external validity in ODA evaluation, unfortunately, has not been popularised. This is partly due to the paucity of attempts by the evaluation community to establish external validity. We have thus proposed a systematic deductive method for establishing external validity. We relied on the Bayesian framework as we believe it to be the most natural choice in analysing the learning processes. But it is by no means the only way, and the suggested directions are shown to be the same from both the Bayesian and frequentist perspectives. We have suggested that the data and analysis should be shared among the evaluators, and that the evaluation community should work more closely with the research community to assist the learning process.

Our conclusion somewhat echoes the conclusions of the empirical growth literature and the aid effectiveness literature. As noted in Durlauf (2003) and Durlauf, Johnson, Temple (2005), the cross-country panel growth regression analysis has its limits in drawing conclusions from data by its model uncertainty. Roodman (2007) cites Leamer's lecture on specification search and concludes by pointing out that the main source of fragility of the cross-country aid-growth literature is the choice of controls.

Thus far, we have pointed out many difficulties of introducing rigorous impact evaluation in the bilateral aid institutions. At the aid institution or organisation level, a set-up for conducting rigorous impact evaluations and accumulating more experiences will be very beneficial for the international aid community. However, a centralised solution, such as creating a new international organisation, is not feasible because of difficulties in both methodology and process. A decentralised approach, starting from new basic guidelines of impact evaluation and communication and learning best practices among the international aid community seems the only solution, at this stage, for rigorous measurement of aid impact. We believe that the issue posed in our article deserves more analytical work by researchers and evaluators in the aid community.

---

<sup>i</sup> JICA and JBIC are planned to be integrated in October 2008.

<sup>ii</sup> For further details, please see the final evaluation report at the JBIC website ([www.jbic.go.jp/english/oec/post/2006/pdf/te03\\_full.pdf](http://www.jbic.go.jp/english/oec/post/2006/pdf/te03_full.pdf))

<sup>iii</sup> There can also be sample selection bias on the observables, but this can be controlled by modelling attrition with them, so we will not regard this as a problem. See Wooldridge (2002: 585–90).

<sup>iv</sup> For further details, please see the final evaluation report at the JBIC website ([www.jbic.go.jp/english/oec/post/2006/pdf/te02\\_02\\_full.pdf](http://www.jbic.go.jp/english/oec/post/2006/pdf/te02_02_full.pdf)).

<sup>v</sup> These are: longitude, latitude, altitude, number of inhabitants, number of dwellings, distance to district capital, distance to national capital, existence of primary school, existence of medical centre, water availability, electricity availability, sewage availability, infant mortality rate, illiteracy rate, qualifies as very poor, qualifies as extremely poor. See Table 5.13 of GRADE (2007).

<sup>vi</sup> See the internal investigative report by the Department of Institutional Integrity of the World Bank (2005).

<sup>vii</sup> We compute  $\Pr[Y]$  by using the likelihood and the prior. Suppose there are two possible models. Then we observe  $Y$  only under the two models, and their probabilities conditioned on each model, or the likelihood of each model, can be

---

written as  $\Pr[Y|M_1]$  and  $\Pr[Y|M_2]$ . So the (total) probability of observing  $Y$  is an average (expected value) of  $\Pr[Y|M_1]$  and  $\Pr[Y|M_2]$ :

$\Pr[Y] = \Pr[Y|M_1] \Pr[M_1] + \Pr[Y|M_2] \Pr[M_2]$ . This depends on the prior  $\Pr[M_i]$ , but it is not problematic because they can be cancelled out once we take the posterior odds ratio  $\Pr[M_i | Y] / \Pr[M_j | Y]$  when we assess the relative relevance of some models  $M_i$  and  $M_j$ .

<sup>viii</sup> There are known implementation issues in BMA which are actively studied: model search, convergence, choice of prior, and comparison over algorithms. Freely available programmes exist to carry out the computation.

<sup>ix</sup> There are studies comparing the relative merits and performances of these methodologies (see references cited in Raftery and Zheng 2003). However, we believe the Bayesian methodology to be more flexible and intuitively appealing.

<sup>x</sup> The frequentist approach is the mainstream school of thought in statistics. Almost all the school curricula up to high school in developed countries teach the frequentist method. Bayesians are the minority school, but this is gaining popularity as computing power becomes cheaper with technological advances in hardware and the development of new computing algorithms.

## References

- Asian Development Bank (2006) *Impact Evaluation: Methodological and Operational Issues*, Manila: Asian Development Bank
- Banerjee, Abhijit V. (ed.) (2007) *Making Aid Work*, Cambridge: MIT Press
- Burnside, C. and Dollar, D. (2000) 'Aid, Policies, and Growth', *American Economic Review* 90.4: 847–68
- Development Assistance Committee, OECD (2007) *Aid Statistics, Donor Aid Charts, Japan*,  
[www.oecd.org/countrylist/0,3349,en\\_2649\\_34447\\_1783495\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/countrylist/0,3349,en_2649_34447_1783495_1_1_1_1,00.html)
- Development Assistance Committee, OECD (2003) *Peer Review, Japan*,  
[www.oecd.org/document/10/0,3343,en\\_2649\\_33721\\_22579914\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/10/0,3343,en_2649_33721_22579914_1_1_1_1,00.html)
- Durlauf, Steven N. (2003) *Policy Evaluation and Empirical Growth Research*, [Working Papers Central Bank of Chile](#) 205, Central Bank of Chile
- Durlauf, Steven N., Johnson, Paul A. and Temple, Jonathan R.W. (2005) 'Growth Econometrics,' in Philippe Aghion and Steven N. Durlauf (eds) *Handbook of Economic Growth*, Volume 1A, Amsterdam: Elsevier
- Easterly, W., Levine, R. and Roodman, D. (2004) 'New Data, New Doubts: A Comment on Burnside and Dollar's "Aid, Policies and Growth (2000)"', *American Economic Review* 94.3: 781–84
- Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B. (1995) *Bayesian Data Analysis*, London: Chapman and Hall
- GRADE (Group for the Analysis of Development) (2007) *Improvement of Living Environment and Livelihoods in Poor Communities in the case of Peru*, JBIC Evaluation Report on ODA Loan Projects FY2006  
[www.jbic.go.jp/english/oec/post/2006/pdf/te02\\_02\\_full.pdf](http://www.jbic.go.jp/english/oec/post/2006/pdf/te02_02_full.pdf)
- Hoeting, Jennifer A., Madigan, David, Raftery, Adrian E. and Volinsky, Chris T. (1999) 'Bayesian Model Averaging: A Tutorial', *Statistical Science* 14.4: 382–417
- Japan Bank for International Cooperation, *Evaluation Report on ODA Loan Projects*, various years
- Ministry of Foreign Affairs, Japan (2003) *ODA Evaluation Guidelines*, Tokyo: Evaluation Division, Economic Cooperation Bureau

- 
- Patton, Michael Quinn (1997) *Utilization-Focused Evaluation: The New Century Text*, (3<sup>rd</sup> edition), Thousand Oaks, CA: Sage Publications
- Raftery, Adrian E. and Yingye Zheng (2003) ‘Discussion: Performance of Bayesian Model Averaging’, *Journal of the American Statistical Association* 98.464: 931–38
- Roodman, D. (2007) ‘The Anarchy of Numbers: Aid, Development, and Cross-Country Empirics’, *The World Bank Economic Review* 21.2: 255–77
- Savedoff, William D., Levine, Ruth and Birdsall, Nancy (2006) *When Will We Ever Learn? Improving Lives Through Impact Evaluation*, Report of the Evaluation Gap Working Group, Washington DC: Center for Global Development
- Todd, Petra E. and Kenneth I. Wolpin (2006) ‘Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility’, *American Economic Review* 96.5: 1384–417
- Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press
- World Bank (2006) *Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank*, Washington DC: World Bank
- World Bank (2005) *Report of Investigation into Reproductive and Child Health I Project Credit N0180, India*, Washington DC: Department of Institutional Integrity, World Bank
- World Bank (1998) *Assessing Aid: What Works, What Doesn’t, and Why*, World Bank Policy Research Report