

## **Review of World Bank Projects**

### **Joshua Angrist, MIT**

My review starts with a short description of each project component, followed by brief narrative reviews of the papers or other documents I read, and a four-category score using the score scale from Annex 3.

Following these thumbnail reviews I make some general comments about the studies I read. I conclude with a response to the questions in Annex 2.

#### **A. Project 75 - Determinants of Schooling**

##### 1. Retrospective vs. Prospective Flip Charts (Glewwe et al)

###### Description

This study contrasts prospective (randomized) and retrospective (observational) evaluation research designs for a school intervention. The intervention is modest (flip charts for use in the classroom), but the lessons of the study are telling. The study is, in effect, a Lalondization of a developing country field evaluation (following Lalonde's influential comparison of randomized and observational evaluations of subsidized training programs). In this case, the observational study shows large benefits of flip charts (arguably, too large to be believed), while the randomized trial shows no effect.

###### Brief narrative review

While the immediate policy question is of limited importance (benefits of flip charts), the lessons from the study are likely to be of general interest. The study provides an important cautionary tale about the consequences of omitted variables bias in observational studies. The paper is well crafted, up to the highest standards of research and scholarship.

##### 2. Estimating Wealth Effects . . . ( Filmer and Pritchett)

###### Description

This is a widely cited study that looks at the relation between household wealth and children's school enrollment in India. It uses principal components methods to construct an index of household wealth and then regresses enrollment on index quintiles and controls. There is also an attempt to deal with measurement error in the wealth variable.

###### Brief narrative review

This paper is a scholarly success in the sense that it is well cited. Moreover, the question of how wealth affects school enrollment is clearly of some scientific interest. The theory of human capital suggests there need not be a strong relation if capital markets function

well. So we'd like to know something about this link. On the other hand, the policy relevance of this relation seems limited. Very likely there is a relation in most data sets. But the identification strategy used here seems unlikely to tell us whether this relation is causal. And even if so, the policy relevance of the findings is unclear. What's the benchmark for the relation, since it seems likely to be ubiquitous? What sort of policy predictions is the extent of the relation input for? Are policymakers in the business of changing wealth per se?

The use of factor analysis to construct an index makes the interpretation of findings especially difficult. This construct is unlikely to be a policy variable in and of itself, or even transparently related to other policy variables. I note, however, that the use of factor analysis in this context gained something of a following. I wish it hadn't. As the authors note in their related JDE paper (study 3, below), the factors themselves are hard to interpret and indeed factors after the first are a complete mystery. The same goal – aggregating indicators – could have been accomplished more transparently by using simple regression imputation in a data set that has both wealth and wealth predictors. We can then decide whether this predictive relation is stable enough to justify use for aggregation of indicators in other samples.

I was also unhappy with the methodological focus in the part of the paper on measurement error – the claim about measuring the relative signal-to-noise ratios of two variables by comparing the relevant IV and OLS estimates. This seems to sidestep a central empirical issue: the expenditure and wealth factors measures used here seem to almost certainly measure different things. Expenditure, for example, is probably more sensitive to transitory income than the components of durables.

### 3. The Effect of Household Wealth . . . (Filmer and Pritchett)

My assessment of this study is essentially the same as study 2 in this project group. While there are indeed some interesting cross-country regularities in the enrollment/wealth relation, the policy relevance of this strikes me as low. The scientific value in this case also seems limited by the clear absence of a causal interpretation for the relation of interest. To the authors' credit, this study notes the difficulty in interpreting the principal components, particularly when more than one component has a large factor loading.

### 4. Worms . . . (Miguel and Kremer)

#### Description

This study is an analysis of a group-randomized trial of a de-worming intervention in Kenyan schools. Intestinal worms are a major public health problem in Africa. The authors look at effects on health and school participation. They leverage the Group-Randomized Trial research design to get at external effects on treated students. The results show large external effects, in addition to substantial benefits for treated students themselves. Previous medical studies had missed this. This study also has an instrumental

variables component converting intention-to-treat effects into causal effects of worm infection rates.

Brief narrative review

This study has both immediate policy relevance (effects of de-worming) and a number of important implications for related public health questions. The study shows how to get the most scientific value out of a well-designed randomized trial. It is well crafted, and the analysis and write-up are up to the highest standards of research and scholarship.

## 5. Economic Openness (Gradstein et al)

Description

This is a theoretical and empirical study of the link between openness to trade and the demand for education. The paper can be seen as falling into (a mostly labor) literature on institutional determinants of the demand for skills. The theory in the paper is in the spirit of work by Acemoglu and coauthors; the empirical evidence comes from a cross-country panel and China.

Brief narrative review

The study is of scientific interest in that it addresses a well-defined causal question about one of the possible effects of trade. On the other hand, the relevant theory has mostly been done already, so the primary contribution should be seen as coming from the empirical side. I did not find the empirical work in the paper very compelling. The identification strategy is not convincing and in some respects the work is technically weak (e.g., panel models were estimated without country and period effects; there was no discussion of the relevant inference issues such as serial correlation). The argument for a demand-side interpretation was not very well substantiated. The reporting style is also not up to standard, with empirical results reported equation-style in the text instead of well-organized tables. The case for policy relevance is mixed. On one hand, openness can be directly affected by policy measures. On the other, a better study would exploit discrete changes in trade policy as a source of possibly exogenous variation in openness.

## **B. Project 9 - Evaluating the Impact of HIV/AIDS Prevention**

### 1. Evaluating the Impact of HIV/AIDS Prevention in Primary Schools (Bundy)

Description

This study describes an AIDS education intervention in Kenya, focusing on children. The study uses a random-assignment research design, involving the training of 540 teachers in 176 schools randomly selected from a pool of 351 schools in western Kenya. The

research plan is to measure the impact of this training on the teachers themselves, and on students. The study is still in the field.

Brief narrative review

This is the sort of policy-oriented micro-development research that I favor. The question is of considerable scientific and immediate policy interest. The evaluation appears to use a well-designed group-randomized trial. The analysis is to be carried out with a team of respected outside scholars, experienced with research of this type, although there are no results or write-up to review as yet.

### **C. Project 82 - Child and Literacy**

#### 1. Is Literacy Shared Within Households . . . (Basu et al)

Description

This study looks at the effect of one adult's literacy on the earnings of others in the same household. The paper has both a theoretical and empirical component.

Brief narrative review

The theoretical section lays out a simple model showing why one's literacy might be worth transferring and what the effects of this on others might be. The paper does not make strong claims for the theory. The question of external effects of literacy is of some scientific interest (as external effects always are) and of some policy relevance. On the other hand, a pre-requisite for credible identification of external effects is credible identification of own-effects. The paper makes no attempt to answer this more basic question. The own-effects provide an essential benchmark and also a test of the identification strategy. The identification of causal effects in this paper relies on a structural sample-selection model. Methods imposing weaker assumptions would have been preferable. Moreover, the exclusion restrictions that lay behind this approach are not justified or substantiated empirically. On balance, this work appears to be behind the state-of-the-art in empirical development.

### **D. Project 90 - Teacher Incentives**

#### 1. Teacher and Principal Incentives in Mexico (McEwan and Sanibanez)

Description

This study evaluates the effects of a point scheme used to improve teacher incentives in Mexico. The evaluation uses a version of a regression-discontinuity design, although there is no sharp discontinuity to exploit.

Brief narrative review

The question of how to improve teacher incentives is of considerable scientific and policy interest in both developing and developed countries. The research design used here is promising, although not quite as strong as would first appear. The authors use an RD framework, but the incentives they study were not assigned as a discontinuous function of the underlying running variable (points). Although not published in a refereed journal, the study is well crafted and the results nicely presented.

## **E. Project 93 - Implementing Affirmative Action**

### 1. Implementing Affirmative Action in Public Service (Zhou Yongmei)

This is a collection of descriptive studies for four developing countries, each produced by a separate author. There is no impact evaluation component to this work, although it might fall under the heading of “process evaluation.” This collection of papers strikes me as being of minimal scientific and policy relevance. Studies like these might be of some help in the formulation of more focused research questions. In my view, however, this sort of work should be given low priority.

## **F. Project 94 - Impact Evaluation of Education Interventions**

### 1. Getting Girls Into School . . . (Filmer and Schady)

Description

This study evaluates a conditional cash transfer program designed to boost school enrollment for 7th grade girls in Cambodia. The intervention used a school-based assignment mechanism. The evaluation design in the study exploits various features of the selection mechanism to construct a regression-discontinuity evaluation. The core estimation strategy, however, is differences in differences.

Brief narrative review

CCT-type programs are of growing interest and importance. The results of this study should therefore be of considerable scientific and policy interest. It’s not clear, however, why girls are of primary interest. In many countries, including many LDCs (like Colombia), girls already get more schooling than boys. In fact, evidence to date on CCTs suggests that boys may be hard to help (e.g., Angrist and Lavy, 2002). On the methodological side, this study has many attractive features. However, the authors miss the opportunity to make full use of a fuzzy-RD/IV strategy based directly on the known discontinuity and nonlinearity in the assignment mechanism (as in, e.g., Angrist and Lavy, 1999, who use a similar rule to estimate the effects class size). This is briefly

considered in Table 5, but more could be done to produce an exceptionally compelling paper. Although not yet published, and in need of further polish, the study is generally well crafted and the results nicely presented.

## **G. Project 167 - Demand for Malaria Vaccine**

### 1. The Demand for Malaria Vaccine: Evidence from Ethiopia (Cropper, et al)

#### Description

This study uses the willingness-to-pay (WTP) method to estimate the demand for malaria vaccine in Ethiopia. The authors argue that WTP estimates of the value of a malaria vaccine exceed cost-of-illness (COI) estimates.

#### Brief narrative review

In the absence of direct evidence on vaccines, WTP may be useful. But it's hard for me to assess the policy relevance of this sort of information. I can imagine that WTP might be quite low, but the case for a public health intervention would still be strong (e.g., a vaccine that has little individual benefit in a population that is mostly vaccinated). On the other hand, WTP might be quite high but the case for a public health intervention still quite low (e.g., a lifestyle drug).

Until such time as there is an actual Malaria vaccine, I would prefer to see impact evaluation of similar public health interventions used as the core input in the evaluation of proposed new interventions. Lessons from polio or MMR might be relevant.

## **H. Project 168 - Which Doctor**

### 1. Which Doctor? Combining Vignettes and Item Response . . . (Das and Hammer)

#### Description

This is a descriptive study that looks at the distribution of doctor quality using the vignette method in Delhi. The authors are especially interested in whether the poor get worse care than the rich, and whether this is true for public as well as private providers.

#### Brief narrative review

I take it as a truism that there is variability in the quality of medical providers, however measured, and that the rich tend to get better care than the poor. It seems likely that this is also true conditional on provider class. As with some of the other studies, I feel I must ask: "what's the benchmark?" I would make the case that the poor get worse everything, and that this is also true conditional on characteristics like eligibility for public care. In

the US, for example, the public schools attended by the children of the rich are better than the public schools attended by the children of the poor. So what? Who needs to know, and what are they going to do with the information? Not knowing the answer to these questions, this study seems to me to be of modest scientific value and limited policy relevance. I would prefer to see research effort directed to the analysis of changes in medical service delivery systems or therapies, focusing on the effects of these changes on precisely-defined patient outcomes.

## **I. Project 169 - Teacher Shocks**

### 1. Teacher Shocks and Student Learning . . . (Das et al)

#### Description

This study uses a unique data set linking student and teacher characteristics in Zambia to study the effects of teacher absenteeism on student achievement. The core econometric method relies on differenced equations that implicitly allow for teacher and student fixed effects in levels.

#### Brief narrative review

This study addresses an issue of major policy interest in developing countries – high rates of teacher absenteeism. The study shows that teacher absences are associated with lower student achievement, even after controlling for various fixed effects. This is a finding of both scientific and policy interest. At the same time, the study misses the opportunity to implement a more convincing research design whereby potentially exogenous shocks to teacher health are used as instruments for absenteeism.

## **J. Project 88 - Health/Education/Employment**

### 1. Poverty, Family Health Problems . . . (Hannum et al)

#### Description

This is a descriptive study of the link between parental health and socioeconomic status on one hand and children's schooling on the other. The study looks at data from three countries.

#### Brief narrative review

This is a long, somewhat unfocused descriptive study. While it provides background information of value, it is not closely linked to questions of either scientific or policy interest. The presentation of results and write-up are up to standard for work of this type.

## 2. Structuring Inequality . . . (Hannum and Adams)

### Description

This study looks at the link between classroom disruption and subsequent school enrollment in China.

### Brief narrative review

The consequences of classroom environment are of some scientific interest. At the same time, it's hard to know what the policy relevance of this study might be. More useful would be an analysis of specific policies designed to improve the classroom environment. In any case, the identification strategy used here is not very compelling.

## 3. Do Health Sector Reforms . . . (Wagstaff and Yu)

### Description

This study reports on an impact evaluation of a large county-based health intervention in China.

The authors focus on a single large province, with variation in treatment arising from the fact that the treatment was not delivered in all counties. The intervention had both an infrastructure and health-insurance component. The outcomes are various utilization and health status measures. The authors note that the program was not implemented with an eye to evaluation. Remarkably, the data-collection component of the intervention studied allowed for baseline data collection only in treated counties and no follow-up.

### Brief narrative review

In an effort to obtain control observations, the authors use a non-program panel data set to track treatment and control counties over time. They use regression, differences-in-differences and a semi-parametric matching/DD hybrid. These procedures appear to be implemented correctly and the resulting output is of both scientific and policy interest. Rather too much attention is given to technical aspects of the matching procedures, however (e.g., use of propensity score) and not enough to the institutional details of treatment assignment and how these details affect evaluation efforts. The value of the study findings for policy is diminished because the results are inconclusive. While this is not the authors' fault, a deeper investigation of the assignment mechanism might help us know which estimates to prefer. On the technical side, it is unclear whether the inference procedures used here with individual and village data make appropriate adjustments for clustering and serial correlation. The most important lesson arising from this work may be that future programs of this sort should include provisions for impact evaluation.

#### 4. Children's Agency . . . (Hannum)

##### Description

This is an essentially descriptive study of schoolchildren in China. It begins with a tabulation of reasons cited for leaving school, as reported by schoolchildren and their mothers. The paper then links school enrollment and academic performance in 2004 to earlier measures of performance and attitudes.

##### Brief narrative review

The first part of the paper is of some value in that it may help pinpoint areas for specific policy intervention. I found the results in the second part difficult to interpret. It seems unlikely these can be taken as reliable measures of specific causal relations. Overall, this study seems to me to be of limited scientific and policy interest. The reporting style and write-up are acceptable.

#### **K. Project 86 - Providing Unemployment Benefits**

##### 1. The Welfare Consequences of Alternative Designs . . . (Hopenhayn and Hatchondo)

##### 2. Unemployment Insurance Savings Accounts: Simulation results for Estonia (Vodipevec and Rejec)

##### Description

These studies explore possible implications of a Feldstein-UISA system. The first is a general welfare analysis of UISA. The second is a simulation study of the consequences of UISA for the Estonian labor market.

##### Brief narrative review

The second study strikes me as the more policy-relevant of the two. Since there are few operating UISA systems available for analysis, there may be some value to data-driven simulations of this sort. On the other hand, there are a few existing programs to study. I would rather see more empirical research along the lines of Kugler's (2000) quantitative evaluation of Colombian UISAs than more simulations, or at least some combination of the two approaches.

##### General comments

The quality and policy relevance of the work I reviewed was variable, running from highly policy related research of exceptional scientific value, to reasonably good studies of modest policy relevance, to studies that seemed to me to be neither very good nor very

relevant. I think it would be relatively easy to identify the lowest-quality or least-useful projects ex ante and take them off the Bank's agenda. In my view, most purely descriptive studies should be off the agenda unless their results can be tightly linked to policy questions or scientific issues of major importance to the Bank's policy work. Impact evaluations without a transparent and compelling identification strategy seem to me to have no place on the agenda either. I give credible impact evaluation of Bank sponsored interventions and related interventions highest priority. I am not against science for its own sake. But in my view, beyond the obvious immediate policy relevance of good impact evaluation, the best general science comes out of this sort of work as well.

My portfolio included some studies that did not purport to be impact evaluations or attempts to get an underlying structural causal relation, but were not descriptive data analysis either. Examples include the WTP study of vaccines, the study of medical provider quality, and a UISA simulation study. While there is clearly a role for this sort of thing, it seems to me that the impact evaluation agenda is much more tightly tied to policy. In some cases, an impact study may provide more clearly valuable scientific input on the issues that these non-impact projects addressed. For example, the best evidence we can produce on the possible benefits of one public health intervention probably comes from the impact of other related interventions. UISAs can also be studied directly in countries where they have been tried. It seems unlikely that the scope for useful impact evaluation "by analogy" has been exhausted. Even interventions with a substantial general equilibrium component can be fruitfully studied in this manner, as work by Duflo and others shows.

I would have liked to see more randomized trials in the portfolio of impact studies. Many of the studies I read deal with schools and education production, an attractive setting for random assignment evaluation. Academic scholarship in modern empirical development (building on a trend in the fields of Labor and Public Finance) has moved strongly in this direction. By now, scholars have demonstrated that useful randomized field trials can be done reasonably cheaply in developing country settings. I was pleased, however, to see a number of well-designed randomized trials in my portfolio. In my view, these studies (e.g., Worms, HIV/AIDS education) should be seen as a model.

The US Department of Education's Institute for Education Sciences provides a compelling example as to how to structure a policy-relevant research agenda; The IES has turned officially and enthusiastically toward specific policy evaluation using credible research designs, especially randomized trials but including quasi-experimental methods. In the absence of random assignment, the IES favors high-quality observational designs like regression-discontinuity methods or natural experiments involving some kind of randomization such as charter-school studies. Two studies in the group I reviewed attempted to use some version of a regression-discontinuity design (McEwan and Sanibanez; Filmer and Schady). These studies mark a promising start, although they do not reflect the full potential of this approach. I was disappointed that the Bank's research agenda lags behind modern empirical Labor Economics, Public Finance, and Development Economics in the use of quasi-experimental variation. For example, I saw

nothing in my portfolio like Duflo's study of the effects of Indonesian school construction.

On the organizational side, I believe the Bank should consider a move to extramural research funding as in US federal agencies like the NIH, NSF, and IES. This too is part of the 2002 Education Sciences Reform Act that led to the creation of the IES. While extramural funding is surely not a panacea, a shift toward extramural research with widely-disseminated RFPs and rotating review panels should promote the separation between program operators and evaluators and increase the pool of qualified scholars that works on Bank projects. Of course, the Bank would remain free to set the agenda for scholars who seek Bank funding. As it stands, the process that currently brings outsiders into the Bank research agenda seems somewhat haphazard. I should note that I saw no clear pattern as to the relative quality of the work by Bank insiders and outside scholars. There was good and bad from both groups. But I did wonder about the process whereby the outside work got funded. In my view, some of this would not have passed muster in a hard-nosed peer review of the sort undertaken at, say, IES or NIH.

Whether or not the extramural route is to be pursued, I think there would likely be a payoff to the use of bureaucratic standards for the sort of impact evaluation that I would like to see required as an essential companion to any program that is fielded partly as pilot from which it is hoped to learn something. One of the studies I reviewed noted that a major health-care program in China was fielded with a limited data collection effort involving treated counties only. This sort of situation seems like it can be easily avoided by bureaucratic fiat. Standards of research quality are easy to specify even for impact evaluation studies not involving random assignment: clear endpoints, defined ex ante; contemporaneous data collection on treated units and plausibly comparable control units at baseline, and in follow-up; intention-to-treat and not treatment received as the key division in comparisons. Studies that fail to meet this standard should be seen as sub-standard.

### **Questionnaire on strengths and weaknesses of the Bank's research**

1. Although the value of Bank-sponsored research has been uneven, Bank researchers have made a number of significant contributions to the study of education and health in developing countries. Some of the work I read contains state-of-the-art empirical work. A significant number of studies, however, appear dated or sub-standard. In some cases, the questions addressed seem inappropriate to a policy-focused agenda or low priority.
2. The topics addressed in Bank-sponsored projects seem hit-or-miss. Many topics are policy relevant, but many are not, as discussed in my review of individual projects above. I don't know anything about the Bank's incentive structure.
3. Awareness of substantive country context does not seem to be an issue.
4. I have not reviewed the Bank's data collection effort.

5. I think the Bank's development agenda would be well served by an increased emphasis on transparent identification strategies applied to the analysis of well-defined causal questions, with less purely theoretical work, and less econometric work with a methodological flavor.

6. I don't know anything about how proposals are reviewed or selected. In some cases, there seems to be an element of serendipity in the process. In my general comments, however, I suggest that the Bank's agenda might be well served by moving to a well-publicized extramural research program in the manner of the NIH, NSF, and, most recently, the IES in the Department of Education.

7. See question 1.