Impact Evaluation for Slum Upgrading Interventions

Erica Field[1]

Michael Kremer[2]

[1] Harvard University, NBER, and BREAD. Email: efield@latte.harvard.edu

[2] Harvard University; Brookings Institution; NBER, and Center for Global Development. Email: mkremer@fas.harvard.edu. Phone: 617 495 9145. Cell: 617.905.9099.

TABLE OF CONTENTS

1 Introduction

The 2003 United Nations Global Report on Human Settlements estimates that 924 million people, or 31.6% of the world's urban population, lived in slums in 2001. Although forecasts are difficult, it is generally agreed that this number could greatly increase in coming years in the absence of strong policy interventions. These trends underscore the importance of slum upgrading strategies for addressing the growing problems of urban poverty.

Upgrading projects focus on providing basic services to improve the well-being of low income communities, including a range of infrastructure interventions frequently undertaken in conjunction with social interventions, such as the regularization of areas with insecure tenure. Other infrastructure improvements include water, sanitation, waste collection, housing, access roads, footpaths, storm drainage, lighting, public telephones, schools, health posts and community centers. Social improvements can include better provision of health and education services, day care, training, and social protection programs. With the projected increases in slum population, the demand for urban upgrading interventions is expected to grow.

Given the trends in urbanization and slum populations, slum upgrading interventions may be an important component of the development process. Investing resources in slum upgrading projects should ideally be based on clear evidence of which specific interventions are more effective. What impact do upgrading projects have on the welfare of the population and how can they be improved to meet the needs of the urban poor? Similarly, policymakers need to understand which specific interventions are more effective than others. These questions can be answered by carrying out appropriate impact evaluation studies. However, because of the many facets of upgrading interventions and the difficulties faced in implementation, evaluating their impact can be complex. Comprehensive evaluation involves focusing on a multitude of potential impacts measured at the community, household and individual levels. This report addresses some of the complexities involved in monitoring and assessing the effect of slum upgrading projects, and provides some recommendations for designing impact evaluations.

In Section 1, we describe the general problem encountered in program evaluations, overview impact evaluation methodologies, and argue there is scope for increasing the role of randomized evaluations. We argue that while all programs should be subject to process evaluations, not all programs should be subject to impact evaluations. In some situations the assumptions underlying quasi-experimental methodologies will be plausibly satisfied, implying quasi-experimental impact evaluations can naturally be applied. In addition, for a subset of policy relevant questions that have been identified as priorities by the Bank, undertaking randomized evaluations would be very useful. In Section 2 we then

summarize existing approaches to slum upgrading impact assessment and offer recommendations for evaluators based on lessons learned from past work. Section 3 gives an overview of upgrading project features and focal areas of impact, and Section 4 details data collection strategies, discusses potential areas of impact that have been overlooked in past impact assessment, and provides recommendations for dealing with these problems in evaluation design and data analysis. Section 5 concludes.

2 The methodology of impact evaluations[3]

2.1 The evaluation problem

Any impact evaluation attempts to answer an essentially *counterfactual question*: how would individuals who participated in the program have fared in the absence of the program? How would those who were not exposed to the program have fared in the presence of the program? The difficulty with these questions is immediate: at a given point in time, an individual is observed to be either exposed or not exposed to the program. Comparing the same individual over time will not, in most cases, give a reliable estimate of the impact the program had on him or her, since many other things may have changed at the same time as the program was introduced. We therefore cannot seek to obtain an estimate of the impact of the program on each individual. All we can hope for is to be able to obtain the average impact of the program on a group of individuals by comparing them to a similar group of individuals who were not exposed to the program. The critical objective of impact evaluation is therefore to establish a credible comparison group, a group of individuals who *in the absence of the program* would have had outcomes similar to those who were exposed to the program. This group should give us an idea of what would have happened to the members of the program group if they had not been exposed, and thus allow us to obtain an estimate of the average impact on the group in question. The concept of establishing a credible comparison group, and hence a counterfactual, is critical to any impact evaluation.

However, in reality it is generally the case that individuals who participated in a program and those who did not are different: programs are placed in specific areas (for example, poorer or richer areas), individuals are screened for participation in the program (for example, on the basis of poverty or on the basis of their motivation), and the decision to participate is often voluntary. For all of these reasons, those who were not exposed to a program are often not a good comparison group for those who were, and any differences between the groups can be attributed to two factors: pre-existing differences (the so-called "selection bias") and the impact of the program. Since we have no reliable way to estimate the size of the selection bias, we typically

[3] This section draws heavily on Duflo and Kremer (2005).

cannot decompose the overall difference into a treatment effect and a bias term. Retrospective studies use non-experimental historical data to estimate the impacts of a given policy or program, but as we will discuss, due to the evaluation problem as discussed in this section, retrospective estimates can be substantially biased if they are unable to construct an appropriate comparison group.

2.2 Overview of methodological approaches to impact evaluation

To solve the evaluation problem, program evaluations typically need to be carefully planned in advance in order to determine which group is a likely control group. A number of different empirical methods have been employed in impact assessment, involving both qualitative and quantitative methodologies. Depending on the context of a particular project, qualitative methodologies can complement or substitute for quantitative approaches. In this sub-section we briefly overview qualitative methodologies, and then discuss quantitative methodologies (including randomized evaluations as well as quasi-experimental methodologies) both in general and in the context of slum upgrading. In the following sub-section we will offer some guidance on context in which randomized evaluations and quasi-experimental analyses may be most useful.

It is worth noting that impact evaluations frequently combine multiple estimation strategies. A good example is a paper by Keare and Parris. Here, the authors conduct a pilot evaluation of four urban shelter projects that received lending assistance from the World Bank from 1972 to 1981. The four projects (in El Salvador, the Philippines, Senegal, and Zambia) focused on the construction and development of higher quality housing units and infrastructure. Efforts included equipping the areas with roads, water, power, sanitation, schools, health clinics, and community facilities. Housing materials loans and assistance in small business development were also provided to participants. The paper summarizes the impacts on the participants and the implementation efficiency, with goals of deriving lessons for the workability of future projects. The authors rely heavily on qualitative discussions and comparisons among the four projects, but their assessment also incorporates quantitative approaches such as income distribution analysis to measure the project's accessibility, hedonic pricing techniques to estimate changes in housing quality, and comparisons of average propensities to consume.

Qualitative methodologies

Qualitative methodologies employed in past impact evaluations can be broken into four broad categories. The *qualitative framework approach* attempts to compare initial ambitions with actual outcomes by defining concepts, describing actors and institutions, mapping dynamics between actors and institutions, and categorizing residents' attitudes and perceptions. The resulting evaluation framework is then applied to available data and case studies (Lee 1998; Otiso

2003; Fiore et. al 2000; Coit 1998). *Qualitative comparisons* identify key outcomes of upgrading projects and implementation issues by comparing case studies of two or more slum upgrading projects (Imparato and Ruster 2003). Comparisons are particularly useful when considering the relevant merits of different slum approaches on specific indicators of interest, such as physical topography (Dundar 2001), land ownership and residential mobility (Bassett 2003), and long-term sustainability (Werlin 1999). Qualitative comparisons have also been used to gauge institutional roles and community cooperation, including residents' roles and relationships in different communities (Lee 1998; Krige, Schur, Sippel 1998).

Accounts based on secondary sources rely primarily on news articles, reports, government documents, and other evaluations to draw conclusions about project impact (Lu 1997; Abramson 1997; Mukhija 2001a; Mukhija 2001b; Verma 2000; Werlin 1999; Yun and Yusof 1991). In contrast, *accounts based on primary sources* present a descriptive story based on information gathered from site visits, fieldwork, focus group, in-depth interviews with participants, technicians, officials, NGO staffs, anecdotal evidence, household surveys of participants' satisfaction and perceptions, and direct observation (Fiore et. al 2000; Bassett 2003; Krige, Schur, Sippel 1998). Both types of accounts are most commonly used to support results from quantitative assessment with anecdotal evidence of community-members' or project coordinators' perceptions.

Quantitative approaches: Randomized and quasi-experimental methodologies

It is useful to divide quantitative approaches to impact assessment into the two categories: experimental approaches (*i.e.* randomized evaluations) and quasi-experimental approaches (*i.e.* using methods such as regressions discontinuity design). *Randomized evaluations* are implemented prospectively, prior to project implementation. Evaluators gather baseline data, assign one or more project components to randomly chosen participant groups (such as individuals, communities, schools or classrooms), and assess the efficacy of the treatment by measuring changes over time in treatment relative to control populations with follow-up data. In the case of randomized evaluations, we can be assured that those who are exposed to the program are no different in expectation than those who are not, and thus, a statistically significant difference between the groups in the outcomes the program was planning to affect can be confidently attributed to the program.

In *quasi-experimental studies,* control populations are generally identified ex-post (that is, retrospectively) based on comparison sites available in the data. Researchers have developed alternative techniques to control for bias in retrospective evaluations as well as possible, and a great deal of progress has been made in non-experimental impact assessment, most notably by labor economists (there are numerous excellent technical and non-technical surveys of these

techniques as well as their value and limitations; see Angrist and Krueger, 1999 and 2001; Card, 1999; and Meyer, 1995).

One strategy to control for bias when using retrospective data is to attempt to find a control group that is as "comparable" as possible to the treatment group, at least along observable dimensions. This can be done by collecting as many covariates as possible and then adjusting the computed differences through a regression, or by "matching" the program and the comparison group through forming a comparison group that is as similar as possible to the program group. One possibility is to predict the probability that a given individual is in the comparison or the treatment group on the basis of all available observable characteristics, and to then form a comparison group by picking people who have the same probability of being treated as those who were actually treated ("*propensity score matching*"). The challenge with this method, as with regression controls, is that it hinges on having identified all the potentially relevant differences between treatment and control groups. In cases where the treatment is assigned on the basis of a variable that is not observed by the researcher (demand for the service, for example), this technique can lead to misleading inferences.

A second strategy to control for bias when using retrospective data is what is often called the "*difference-in-difference*" technique: when a good argument can be made that the outcome would not have had differential trends in regions that received the program if the program had not been put in place, it is possible to compare the *growth* in the variables of interest between program and non-program regions. However, it is important not to take this assumption for granted. This identification assumption cannot be tested, and to even ascertain its plausibility one needs to have long time series of data from before the program was implemented in order to be able to compare trends over long enough periods. One also needs to make sure that no other program was implemented at the same time, which is often not the case. Finally, when drawing inferences one must take into account that regions are often affected by time persistent shocks that may look like "program effects." Bertrand, Duflo and Mullainathan (2004) found that difference-in-difference estimations (as commonly performed) can severely bias standard errors: the researchers randomly generated placebo laws and found that with about twenty years of data, difference-in-difference estimates found an "effect" significant at the 5% level for up to 45% of the placebo laws.

Finally, a third strategy to control for bias when using retrospective data, called "*regression discontinuity design*" (see Campbell, 1969), takes advantage of the fact that program rules sometimes generate discontinuities that can be used to identify the effect of the program by comparing those who made it to those who "almost made it." That is, if resources are allocated on the basis of a certain number of points, it is possible to compare those just above to those just below the threshold. Angrist and Lavy (1999) use this technique to evaluate the impact of class size in

Israel. In Israel, a second teacher is allocated every time the class size grows above 40. This policy generates discontinuities in class size when the enrollment in a grade grows from 40 to 41 (as class size changes from one class of 40 to one class each of size 20 and 21), 80 to 81, etc. Angrist and Lavy compared test scores in classes just above and just below this threshold, and found that those just above the threshold have significantly higher test score than those just below, which can confidently be attributed to the class size since it is very difficult to imagine that schools on both sides of the threshold have any other systematic differences. Such discontinuities in program rules, when enforced, are thus sources of identification. However, such discontinuities are often *not* implemented, especially in developing countries. For example, researchers attempted to use as a source of identification the discontinuity in a Grameen bank (the flagship microcredit organization in Bangladesh) policy, which lends only to people who own less than one acre of land (Pitt and Khandker, 1998). However, it turns out that *in practice*, Grameen bank lends to many people who own more than one acre of land, and that there is no discontinuity in the probability for borrowing at the threshold (Morduch, 1998). In developing countries, it is likely to often be the case that rules are not enforced strictly enough to generate discontinuities that can be used for identification purposes.

In the context of slum upgrading, non-experimental evaluations typically involve bivariate and multivariate regression analysis to test hypotheses of household and community responses to infrastructure improvements.[4] For instance, using data from household interviews and direct observation of infrastructure, evaluators can perform regression analysis in which household income serves as the dependent variable and infrastructure type is included in the set of regressors (Aiga and Umenai 2002). With sufficient control variables, the coefficient estimate on the indicator of infrastructure can be used to estimate the effect of specific slum upgrading interventions. In evaluations of housing markets, quantitative methodologies frequently involve tests of the hedonic price model. Here, regression analysis is used to estimate a household's willingness to pay for a bundle of housing attributes (marginal hedonic prices). It is then possible to evaluate a housing subsidy project by deriving direct Hicksian welfare benefits of subsidies for particular housing amenities (Kaufmann and Quiqley 1987; Crane, Daniere, and Harwood 1997).

[4] Quantitative methodologies can also be used to derive predictions of slum upgrading impact prior to intervention through policy simulation and financial analysis exercises. For instance, before a nation-wide expansion, program officers may want to know which among a set of proposed interventions will have the most significant impact and highest probability of financial sustainability in different parts of the country. Policy simulation applies parameter estimates obtained from regression analysis to simulate outcomes among a given population and thereby assess potential changes in welfare under alternative slum interventions (Kapoor et. al 2004). Financial analysis assesses the financial and management capacity of the sponsoring organizations using data from local governments such as cash flow patterns, budgets, financial statements, and progress reports (Krige et al. 1998).

Comparing randomized and quasi-experimental methodologies

The main issue that arises with non-experimental methodologies is that they may contain large and unknown biases resulting from specification errors. A growing literature is taking advantage of interventions which were randomly implemented to estimate the program's impact using experimental methods and then re-estimate the program's impact using one or several different non-experimental methods - thus providing a test of whether the non-experimental estimates are biased. LaLonde's seminal study (1986) found that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates, and that such econometric estimates often differ significantly from experimental results. A number of studies have conducted such analysis with a focus on the performance of propensity score matching, with mixed results. More comprehensive is a recent review study by Glazerman, Levy, and Meyers (2003), who assessed both prospective (experimental) and retrospective (non-experimental) methods in studies of welfare, job training, and employment service programs in the United States, synthesizing the results of 12 design replication studies. Glazerman et al. found that retrospective estimators often produce results dramatically different from randomized evaluations, that the estimated bias is often large, and that they were unable to identify any strategy that could consistently remove bias and still answer a well-defined question.

We are not aware of any systematic review of similar studies in developing countries, but a number of comparative studies have been conducted - some of which suggest omitted variables bias is a significant problem, others which suggest non-experimental estimators may perform well in certain contexts. Buddlemeyer and Skoufias (2003) and Diaz et al. (2003) both focus on PROGRESA, a poverty alleviation program implemented with a randomized design in Mexico in the late 1990s. Buddlemeyer and Skoufias (2003) use randomized evaluation results as a benchmark to examine the performance of regression discontinuity design and find the performance of regression discontinuity design in this case to be good, suggesting that if policy discontinuities are rigorously enforced that regression discontinuity design frameworks can be very useful in some contexts. Diaz et al. (2003) compare experimental estimates to propensity score matching estimates, again using the PROGRESA data. Their results suggest that propensity score matching does well for outcomes measured using similar survey questionnaires, but that for outcomes measured using different survey instruments there are substantial differences. Given that in many retrospective studies survey data may be compiled from a variety of sources, this leaves substantial room for concern; in addition, even when survey data comes from a single source they may often be many fewer variables available than are available in the PROGRESA survey.

Several studies in Kenya have provided evidence that estimates from prospective randomized evaluations can often be quite different from estimated effects in a retrospective

framework, suggesting that omitted variable bias is a serious concern. Glewwe et al. (2004) study an NGO program which randomly provided educational flip charts to primary schools in Western Kenya. Their analysis suggests that retrospective estimates seriously overestimate the charts' impact on student test scores; a difference-in-difference approach reduced but did not eliminate this problem, and it is not clear that such a difference-in-difference approach has general applicability. Similar disparities between retrospective and prospective randomized estimates arise in studies of the impact of de-worming in Kenya (Miguel and Kremer 2004). Future research along these lines would be valuable, since comparative studies can be used to assess the size and prevalence of biases in retrospective estimates. However, when the comparison group for the retrospective portions of these comparative studies is selected ex post, the evaluator may be able to pick from a variety of plausible comparison groups, some of which may have results that match experimental estimates and some of which may not. To address these concerns, future researchers should conduct retrospective evaluations before the results of randomized evaluations are released or conduct blind retrospective evaluations without knowledge of the results of randomized evaluations or other retrospective studies.

2.3 The role for randomized evaluations

Identification issues with non-randomized evaluation methods must be tackled with extreme care because they are less transparent and more subject to divergence of opinion than are issues with randomized evaluations. Moreover, the differences between good and bad non-randomized evaluations are difficult to communicate, especially to policy makers, because of all the caveats that must accompany the results. These caveats may never be provided to policy makers, and even if the caveats are provided they may be ignored: in either case, policy makers are likely to be radically mislead. This suggests that while non-randomized evaluations will continue to be needed, there should be a commitment to conduct randomized evaluations where possible.

It is worth clarifying that we are *not* proposing that all projects be subject to randomized evaluations. Historically, prospective randomized evaluations of development programs have constituted a tiny fraction of all development evaluations. While we argue there is scope for considerably expanding the use of randomized evaluations, these must necessarily remain a small fraction of all evaluations.

All programs should be subject to a "process evaluation" – that is, resource flows should be monitored, and there should be a very basic attempt to gather information on what did and did not go well with the program, in an attempt to learn lessons which can be applied to future work.

Impact evaluations should not be applied to all projects. However, each quasi-experimental methodology rests on a particular set of assumptions, and cases will arise in which these assumptions will be plausibly satisfied – in which case undertaking a quasi-experimental impact evaluation would be worthwhile. Cases will arise (for example) for which a regression discontinuity design could naturally be applied; namely, for situations in which resources are allocated on the basis of a certain number of points, in which case it may be possible to compare individuals just above to those just below the threshold. Such non-randomized methodologies may well be usefully applied in a number of contexts, and should be done so when situations applied in which they can be credibly used. Selectively applying randomized evaluations will be useful as a third option in certain contexts.

The Appendix to this paper draws on previous fieldwork by the authors and others, and argues that randomized evaluations are feasible and can be successfully implemented in a much wider variety of contexts than those in which they have been used in the past. For example, there are many different ways to introduce randomization into an evaluation or research program; which one is appropriate for a given program or research question will depend on the context and programmatic and other constraints, but researchers can draw from a variety of options including setting up an experiment, evaluating a lottery, evaluating a randomized rotation of a program or policy, *etc.* For example, before a program is launched on a large scale, a pilot project (necessarily limited in scope) is often implemented to test the program's feasibility and sometimes to compare the effectiveness of alternative versions; randomly choosing the beneficiaries of the pilot can be done in most circumstances, since many potential sites (or individuals) are as deserving as others to host the pilot. By randomly choosing sites that will receive the pilot and collecting data on these and randomly selected control sites, it is possible to get accurate information on the impact of the pilot program. PROGRESA, a conditional cash transfer program in Mexico) provides an example of such a randomized pilot program. In 1998, when the program was launched, officials in the Mexican government made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential beneficiary communities of PROGRESA all at once, and instead started with a pilot program in 506 communities. Half of those were randomly selected to receive the program, and baseline and subsequent data were collected in the remaining communities.

We would suggest that the Bank identify a set of questions of substantial policy importance; for these questions, it would be very useful to implement randomized evaluations. We will return to this point in the discussion specific to slum upgrading, when we suggest several potential policies which may be good candidates for randomized evaluations in the context of slum upgrading.

Rigorous and systemic evaluations have the potential to leverage the impact of international organizations well beyond simply their ability to finance programs. Credible impact evaluations are international public goods: the benefits of knowing that a program does or does not work extend well beyond the organization or the country implementing the program. Programs that have been shown to be successful can be adapted for use in other countries and scaled up within countries, while unsuccessful programs can be abandoned. Through promoting, encouraging, and financing rigorous evaluations (such as credible randomized evaluations) of the programs they support, as well as of programs supported by others, the Bank and other international organizations can provide guidance to the international organizations themselves, as well as other donors, governments, and NGOs in the ongoing search for successful programs. Moreover, by credibly establishing which programs work and which do not, the international agencies can counteract skepticism about the possibility of spending aid effectively and build long-term support for development.

3 Objectives of impact evaluation of slum upgrading projects

Section 2 discussed the general evaluation problem, and discussed the potential for an expanded role for carefully targeted randomized evaluations. We now shift our focus to impact evaluation issues particular to slum upgrading.

3.1 Overview of project features

Slum upgrading consists of physical, social, economic, organizational and environmental improvements within neighborhoods. These projects may be undertaken by citizens, community groups, businesses and local and national authorities. Typical actions include:

- Regularizing security of tenure through property mapping, titling and registration
- Installing or improving basic infrastructure, including water, waste collection, storm drainage, electricity, security lighting, and public telephones
- Removal or mitigation of environmental hazards
- Providing incentives for community management and maintenance
- Constructing or rehabilitating community facilities such as nurseries, health posts, community centers
- Home improvement, including material upgrading, new construction and expansion of existing structures
- Improving access to health care and education as well as social support programs to address community issues such as crime and substance abuse
- Enhancement of income-earning opportunities through training and micro-credit
- Crime control

Many slum upgrading interventions bundle these services and provisions, so it is critical for evaluators to have a clear understanding of each project component. Furthermore, in many cases subprojects grow out of initial interventions, so evaluators must monitor closely the expansion of services and program components throughout the project life. To do so, evaluators should conduct in-depth interviews with central and field-level project administrators at the start of the process as opposed to relying on intervention blueprints. When data is collected from participants, survey questionnaires should always include open-ended questions about changes in the community infrastructure and social programs operating in the neighborhood.

3.2 Focal areas of impact

Because slum upgrading projects frequently comprise more than one of the above interventions, comprehensive impact evaluations require monitoring a wide range of social and economic indicators spanning many potential areas of impact. Key outcomes monitored in the past can be divided into three categories of evaluation: *Direct program impact assessment* analyzes the success of physical implementation of the project and population access to project resources and infrastructure through indicators such as rates of project completion, rates of coverage and usage, and project maintenance. Where applicable, the effectiveness of targeting is also measured by calculating rates of access, quality, and affordability of relevant infrastructure services to specific socio-economic or demographic groups.

Socio-economic impact assessment measures the impact of slum upgrading projects on both individual and community-level outcomes. Relevant individual indicators include health and schooling attainment, employment, and statements of tenure security. Community-level outcomes include land and housing values, aggregate indicators of poverty, inequality and local economic development, crime rates, environmental risks, local resource management, local institutional development, and integration with government agencies. Finally, comprehensive impact assessment also considers *indirect program effects* at both the individual and community level, such as migration, political enfranchisement, social capital accumulation and the development of complementary infrastructure.

To calculate the cost effectiveness, net value, and sustainability of slum upgrading interventions, impact assessment must also monitor all potential program costs by measuring the rate of cost recovery and the incidence of adverse program effects. This involves close attention to individual and community out-of-pocket payments and any changes in tax revenues accompanying the upgrading effort, in addition to potential adverse consequences of the program on both local and non-beneficiary populations, such as residential displacement, strain on local institutions, in-migration, and crowding out of local providers.

3.3 Specific outcome measures

Here we outline standard approaches to evaluating specific project outcomes. The following broad areas of impact are relevant for the majority of slum upgrading interventions. In general, comprehensive impact evaluations should cover each of these bases at some level. Specific focal outcomes within these categories should be chosen based on the details of the intervention and anticipated effects.

Completion rates

To assess the success of physical implementation, evaluations most commonly gather data on the completion rates of specific project components. This simple measure is important to monitor and frequently overlooked. For instance, while Keare and Paris conclude that, although the programs were "remarkably successful" in terms of increasing housing stock and keeping housing and services affordable and accessible to low-income groups, significant problems occurred with delays in implementation. By the early 1980s, only three of the four slum upgrading projects had completed the majority of infrastructure and housing construction.

Delays in completion of housing construction and occupancy frequently arise on account of slow land acquisition and installation of basic services from lack of coordination among agencies, inconvenient distances, and inadequate credit. Because of interest, inflation, and overlapping house payments, home owners' costs can escalate with delays, reducing the benefits of the project even further. To keep track of delays, evaluators should monitor project completion step by step through ongoing or retrospective data collection. This also serves to help identify specific bottlenecks in implementation in order to reduce costs and hold-ups in future projects. Project completion assessment data can be gathered from household surveys on usage rates, household history of usage (date of access to new services), and out-of-pocket payments (including labor costs) for infrastructure use. In addition, as part of the follow-up survey, it is useful to gather information on community members' subjective assessment of the quality of project implementation and general experiences with new infrastructure and services. Finally, details of project roll-out should be gathered directly from program administrators in a follow-up survey.

Equity of impact

The project's affordability, accessibility, and fairness of neighborhood selection can similarly be gauged from survey respondents' reported experiences with project implementation gathered in household surveys. This aspect of impact evaluation is particularly pertinent when projects were designed with an intention of targeting interventions to the neediest residents. In these cases, evaluations should examine whether projects were affordable to low-income

populations, whether neighborhood selection was fair, and whether specific features of housing or infrastructure improvements were attractive to the poor.

To account for these aspects of project success, three evaluation strategies are commonly employed.[5] Income decile analysis calculates the number of participants that fell within a target income range, and determines what fraction of families in each decile participated in the project and the lowest percentile that could afford to participate. Turnover analysis compares income profiles of dropout families, with special attention paid to the degree of turnover due to inability to pay. Expenditure analysis analyzes monthly housing costs and percentage of income spent on housing, which can then be used to assess ability to pay under alternative scenarios.

In their analysis, Keare and Parris use income distribution analysis and find that projects were generally successful in reaching the target groups. Three important lessons can be drawn from their study. First, in Senegal, problems with affordability arose on account of the fact that incomes did not keep pace with increased costs due to delays in project implementation. This example highlights the importance of collecting direct measures of project implementation (dates of completion) as well as time trends altering the quality of services as a means of identifying specific sources of project failure. Second, in exploring technical issues with affordability and income measurement, the authors note the importance of alternative income sources that can be used to pay for services. For instance, complimentary income-generating activities such as rental income can provide sources of funding for participants and help maintain project affordability. Finally, in analyses of affordability and equity it is useful to consider potential tradeoffs between accessibility and targeting a low but narrow income range. To the extent one believes that there are beneficial effects of mixed income neighborhoods, there might potentially also be a tension between targeting the poor and attracting some better off people to the neighborhood.

Socio-economic impact

Socioeconomic impact evaluation is concerned with evaluating the effect of project participation on individual outcomes such as entrepreneurial income, labor force participation, health, education and employment. The most common methodologies are longitudinal impact studies using experimental or quasi-experimental designs. These evaluations compare changes in per capita expenditures on housing, food, medicine, education and transportation. It is also potentially informative to measure rates of participation in social activities. A frequently overlooked aspect of slum quality that takes into account potential in-migration are analyses of living density and living space, that measure changes in residences' area per person and people

[5] Keare and Jimenez (1983) discuss affordability in greater detail.

per room. Information gleaned from asking participants about their satisfaction with and changes in socioeconomic conditions can also be used in both quantitative and qualitative analyses.

Labor market activity is also a potentially important area of impact. Keare and Parris note that slum upgrading projects may be associated with increased labor force participation rates. Field (2003) finds empirical evidence of such effects in slums of Peru following land regularization. In this example, increases in employment in the range of 10-15% of baseline employment hours arose as a result of increases in tenure security and the accompanying reduction in household members' time spent protecting homes and communities. Field also found evidence of reductions in child labor force participation as adult labor became less constrained, as well as high relocation rates of entrepreneurial activities.

Community-level outcomes

Community-level evaluations track details of neighborhood public goods and services, as well as household participation in cooperatives and neighborhood organizations. Neighborhood public goods provision may be affected by the upgrading project, and the level of community organization may also be an important determinant of project success. Here it is useful to distinguish between types of organizations, differentiating producers' associations, women's groups, political groups, credit associations, and local governance.

Some public goods can be measured by survey-takers when they visit the neighborhood, such as level of cleanliness and feelings of safety. However, the stock of many neighborhood public goods and services may not be sensitive to upgrading interventions in the short run. Hence, it is more practical for evaluators to examine changes in individual participation in local organizations and activities. For instance, evaluators can track households' time and money spent participating in community groups, as well as the number and nature of all community groups at the neighborhood level. Surveys also commonly track the community-members' perceptions of local leadership, including the manner in which local leaders are chosen and the control leaders exercise over local decision-making and resource allocation. In quantitative assessments, it is sometimes useful though uncommon to estimate the opportunity cost of participating in community projects by asking families about foregone income, wages of paid labor, and hours of unpaid labor. It is also relevant to consider gender differences in project participation.

Real estate market effects

Evaluations typically monitor project impact on both quality and value of housing. To do so, standard assessments compare changes over time in the quality, services, neighborhood traits, and housing costs within categories of alternative housing options. It is common to

summarize these outcomes by constructing housing quality indicators based on quality scores for housing attributes and building materials across different types of units. Similarly, summary measures of housing cost used to assess changes in home values can be based on cost of construction or purchase, owner's estimated sale price or rent, and observed market rents.

Housing value data is particularly important for evaluating the impact of slum regularization or land titling programs. When collecting sale and rental price data among titled and untitled residents, one approach is to ask respondents to estimate the hypothetical sales and rental value of their residence under alternative states of tenure status (with and without a title). This allows the evaluator to construct a within-household estimator of program impact on housing value. Whenever collecting data on hypothetical selling or rental values, some argue it is important to clearly specify and hold constant the buyer under different scenarios, or else it is complicated to separate out changes in sales prices due to changes in buyer characteristics from changes in unit value (see Lanjouw and Levy, 2003, for a discussion). On the other hand, one might imagine that values could change if the program affects the set of potential buyers. Evaluations of changes in housing value should control for lot size, living area, materials and construction quality, and access to and delivery of services.

Although the value of housing is an outcome of central interest to many impact evaluations, it is difficult to rely on questions about how much is the house worth for sale and rent, particularly under hypothetical scenarios. It is therefore valuable to verify sales prices with sales documents from matched properties. Survey-takers should be instructed to ask respondents to see a copy of the purchase agreement. In some cases, these transactions may be recorded in local registries. However, it also makes sense to find out if there are typically off-the-record payments in such transactions to evade taxes.

Another potentially valuable approach is to ask local real estate agents for price data on housing units. Real estate agents may also be able to collect more accurate data by asking residents about sales value since they would be perceived as potentially initiating sales and rental transactions in the future. Hiring official housing appraisers to assess unit values may also be possible in some settings. To assess housing improvements and value added with expert assessments, one practical strategy is to collect photographs of the housing units at the time of baseline and follow-up surveys, then ask an architect or engineering advisor to assess the extent of home improvements. While compromising somewhat on quality relative to hiring experts for field assessment, this approach is considerably more practical when skilled labor is scarce.

When collecting data on the amount invested in housing improvements directly from homeowners, rather than asking respondents to report the value of investments, it is generally more reliable to ask respondents to report the amount and type of material used in construction and then to calculate investment amounts based on local prices and labor services collected from local construction organizations. Finally, in addition to tracking actual housing investments and property transactions in local real estate markets, evaluators may want to ask participants about *intentions* to invest in, rent or sell property in the future (specifying an exact reference period). This is particularly useful in the case that follow-up surveys are conducted shortly after the intervention is finished, in which case too little time may have elapsed to observe important program effects.

Housing data can be used to conduct many types of quantitative analyses. For instance, evaluators may want to characterize general trends in the area's rental or sales market, analyze returns to investment on housing or land, compute future income flows from rental markets over 20-30 year horizons, or estimate the benefits and return on specific housing investments or changes in ownership security (see Kaufmann and Quigley, 1987; Lanjouw and Levy, 2003).

Cost recovery

Another basic outcome of interest is the financial sustainability of the project. To estimate the degree of cost recovery associated with a given intervention, evaluators should consider the number of participants that completed payments on time, the degree to which payments were delayed and the amount owed. In addition, evaluators should assess the reasons for unexpected problems with cost recovery by tracking changes over time in program costs, income levels, or cost of living. In their analysis of housing projects in four countries, Keare and Parris note that community involvement was important in enforcing repayments and facilitating housing relocations when necessary.

3.4 Common evaluation issues for urban upgrading projects

Past evaluations have encountered challenges in the following areas of impact assessment.

Sustainability

It has frequently been difficult in past evaluations to accurately evaluate the sustainability of slum upgrading interventions. This is important since pre-project cost-benefit analyses typically rely on relatively long time horizons for cost recovery or projected welfare benefits. When evaluations take place relatively soon after projects are completed, it is very difficult to discuss long-term effects. Because of the importance of this aspect of project welfare benefits, evaluators should urge managers to follow intervention populations at regular intervals at least

five years after project completion. For instance, the optimal evaluation design that minimizes costs may be to follow project participants one year and five years after project completion.

Without long-run follow-up data, evaluators have relied on projections of sustainability based on short-run project costs, completion and usage rates. In reality, predicting the sustainability of projects requires more careful consideration of the vulnerability of the project to economic and political shocks. This is an inevitably complicated projection formula. In some cases problems of sustainability may arise very quickly after the project is completed, providing a good indication of future issues the community will face in maintaining infrastructure. For instance, Keare and Parris noted in their evaluation that problems were already emerging with garbage collection in Zambia. However, since the projects were completed fewer than two years before the report, the authors had inadequate evidence to assess the likelihood of program continuation.

The best way to assess sustainability is through an analysis of project participation and implementation through changes in political regimes and economic shocks. As opposed to confounding impact evaluations, economic or political changes can provide important information about the likelihood that interventions will be sustainable in the future. Evaluators should look for such opportunities to shed light on these projections.

Complementary programs

A critical component of the benefits of slum interventions is the degree to which complementary services and programs arise or respond to community upgrading. For instance, the value of road building depends heavily on the extent to which bus services proceed to enter the neighborhoods. In this case, the rate or frequency of community members' use of roads is a poor indicator of the welfare benefit of the intervention if it does not take into account changes in available forms of transportation. Here, the relevant outcome measure in a program evaluation would relate directly to complementary goods and services brought into the neighborhood as a result of upgrading. Similarly, homeowners may not be in a position to respond to increases in tenure security through housing investment if they are sufficiently credit constrained. In these cases, the value of urban upgrading may rely on complimentary private or public interventions such as housing materials banks or micro-credit programs.

Careful impact assessment should identify not only the net effect of the program but also the specific catalysts of failure or success. To do so, impact evaluations need to keep careful track of all potentially important complimentary services that would significantly raise the value of the intervention. This is important both for evaluating benefits (as in the case of roads), and

for identifying barriers to project success. Such analysis can generally be done prospectively, based on careful consideration of anticipated outcomes. However, assessment of these channels must then be incorporated into data collection efforts and participants' subjective assessment of project success.

Administrative costs

Upgrading projects have failed more than once in the past due to unanticipated costs of administering the program. For instance, a large materials loan program in Zimbabwe ran into significant administrative difficulties and costs, captured in the initial evaluation. In this case, the loans had been kept small to avoid burdening families, but as a result were frequently insufficient to finance construction. Many evaluations fail to keep track of administrative difficulties and costs in follow-up surveys and monitoring efforts, relying on pre-intervention projections of management costs to evaluate the net impact of the program. To avoid this type of estimation error, and to identify problems with implementation that could be avoided in the future (as was the case in Zimbabwe), evaluators should keep track of project implementation costs over time by interviewing project coordinators and administrators throughout the project cycle.

Quality control

Past evaluations have proven that it is difficult and costly to monitor infrastructure quality. This complicates analyses of project completion rates, as well as changes in housing values and attributes and investment incentives. In the past, evaluators have frequently relied on self-reported data from households and project administrators. More accurate data can be collected by hiring engineers and architects to gather comprehensive information on the quality of constructions and specific inputs used in project-related infrastructure and community investments (Olken, 2005). When local experts are unavailable, construction quality can be assessed indirectly by asking households or community groups to report the amount and types of inputs used in construction, although this does not circumvent the potential problem of imperfect recall and intentional misreporting in the case that corrupt project administrators skim project funds by skimping on materials.

Local participation

It is unclear from past studies the extent to which successful project implementation relies on local participation. Many policy-makers currently advocate designing poverty interventions from the bottom up, including working with communities at each stage of project design and letting communities decide what levels of service they receive. Similarly, many past evaluations of slum upgrading projects have concluded that interventions are most effective when led by the municipal authority and implemented at the community level through a broad

set of intermediaries including community based organizations, NGO's, and UN agencies such as UNICEF and Habitat. Evaluators can potentially shed light on these claims by designing evaluations that randomize the level of community participation and integration. Given that this is an important aspect of project design that has not been thoroughly examined, it would be highly beneficial for a range of poverty interventions to test these hypotheses through a carefully designed experimental study.

4 Recommendations for comprehensive impact assessment

4.1 Data collection strategies

Although sources of data used for impact assessment may include case studies, interviews with participants, staff, and organizations, direct observations, and secondary sources, socioeconomic household surveys (longitudinal and cross-sectional) provide the basis of most evaluations. In conjunction with household surveys, it is frequently useful to consult experts to evaluate project features such as construction quality.

Panel versus cross-section data

To carefully monitor the impact of any poverty intervention, it is necessary to collect panel data. Ideally, the baseline study should be completed before the intervention begins, and followed up at regular intervals following the project's completion. To the extent that the announcement of a project will affect land markets and potentially residential choice and housing investment, baselines will ideally be conducted prior to the announcement of which areas will be covered by the project. Because attrition and time trends confound data analysis, follow-up surveys should begin shortly after the project is completed. However, since the most important effects of the program may be long-term, it is wise to follow households at least three years post-intervention. The timing of the follow-up survey should be determined by the setting and anticipated impacts, and based on community-level data such as migration rates and economic and political volatility.

To understand the impact of the program on neighborhoods, it is a good idea to interview new residents in homes where original respondents have moved. Even if complete data is not collected for such households, data on infrastructure, purchase prices, and housing quality can be gathered from current residents. This will also allow a short-term analysis of changes in neighborhood composition accompanying residential mobility.

To understand the impact of the program on people, it will be key, however, to establish a baseline of residents in the treatment and comparison areas and follow these up over time, whether or not they remain in the project area or control neighborhood. Following all original sample members is important since it is likely that the group of individuals who leave the neighborhood is not a random sample of community members, such that a pre-post comparison of remaining individuals will not accurately reflect the impact of the program on the entire population. For instance, residents who are excluded from access to new infrastructure may be disproportionately likely to leave the community, which would lead to an overestimate of project accessibility rates. For similar reasons, it is also important to follow all original sample members from control communities. In particular, it is likely that mobility rates will differ between project and non-project areas. If attrition patterns differ according to whether sample members live in control or treatment regions, difference-in-difference estimates of socio-economic impacts of the project will be biased.

Arguably there is no point in doing an impact evaluation without a baseline because there is no way to know who should have been considered in the treatment and comparison groups. However, if for some reason it is necessary to do an impact evaluation ex post, cross-section data can be used in an attempt to approximate longitudinal data with retrospective questions. It is important to be cautious in using retrospective survey data to construct pseudo-panel estimates because of potential recall bias. To minimize this concern, evaluators may want to anchor questions around specific events to improve recall. For this purpose, the project itself is a particularly useful means of anchoring. For instance, household surveys can include a module related to program impact in which participants are asked to provide information about employment, health, time use, household composition and community activities before and after the program. In this module, questions can be asked both of perceptions and actual impacts.

Pre-baseline surveys

In many cases, qualitative work is needed to figure out what quantitative questions to ask. New slum upgrading interventions may benefit substantially from conducting "pilot surveys" that are largely qualitative in nature in order to assess the most important areas of impact. This would typically involve survey-takers spending a few days in communities in which projects will be introduced, and discussing with residents and project administrators the anticipated benefits of the program. This is also a useful method for determining potential bottlenecks in project implementation that should be monitored with care, and getting a sense of local issues of importance to particular communities.

Community level questionnaires

In addition to using household questionnaires, evaluators should also conduct community-level questionnaires that collect data on neighborhood resources, organization and history from local leaders, long-term residents, and administrators of local organizations. In these surveys, it is particularly important to find out what other interventions are occurring in the community. For questions from community leaders, survey-takers should identify all potential local leaders from different neighborhood positions (such as church leaders, political leaders, and cooperative association leaders).

Sample design[6]

Most evaluation methodologies will involve household survey data collection, which raises a number of sampling issues particular to urban slums. One potential complication is that poor urban neighborhoods are frequently underrepresented in census data, particularly in the case of illegal settlements or settlements that the government does not want to formally recognize. In these cases, to identify an appropriate sampling frame, evaluators may need to collect data on the universes of intervention households from local authorities. There are many potential sources of administrative data on slum populations from which a sample of target households can be randomly selected. For instance, local cadastres or municipalities frequently compile geo-spatial data on housing units for planning purposes. Similarly, address data may be available from electoral rolls – although these data may face the same limitations of national census data in terms of under-representation.

Collecting adequate survey data for nationwide or multi-site upgrading interventions usually requires a two-stage sampling procedure, in which communities are randomly selected from the universe of urban neighborhoods and then clusters of households are selected from within the chosen neighborhoods. Even when a complete roster of households is available from the national census, a frequent challenge in conducting two-stage sampling is identifying the appropriate sampling frame for first-stage selection of neighborhoods. In many cases, the relevant neighborhoods or communities targeted for intervention are not well-defined geographical units that can be easily mapped to census data. Particularly in the case of urban

[6] It is not possible to adequately cover all of the issues relevant to the design of impact evaluations in this paper, nor to cover at a deep level the details involved with (for example) sample sizes, subgroup analysis, attrition, *etc.* Deaton's *The Analysis of Household Surveys* (1997) provides an excellent overview of many issues relevant to survey design. Meinart's Clinical *Trials: Design, Conduct, and Analysis* (1986) is a standard reference for the design of randomized trials in particular. Esther Duflo, Rachel Glennerster, and Michael Kremer will provide a discussion of these issues in the context of randomized evaluations of development programs in a *Handbook* chapter currently in preparation.

informality, geographic boundaries of treatment areas may be inconsistent with size- or population-based geographic units.

In this case, one option is to choose the smallest exhaustive urban geographic unit available in the census data (such as a census segment or block), and then use population-weighted sampling techniques to select a random set of urban locations. Once the blocks are chosen, evaluators can go on-site to locate the neighborhoods boundaries and aggregate neighboring census blocks into treatment units. At the same time, evaluators can correct incomplete household rosters within the identified set of communities. Indeed, this is one of the major advantages of multi-stage sampling: Since the list of members is required only for those clusters used in the final stage, multi-stage sampling does not require a complete list of members in the target population, which greatly reduces sample preparation cost.

A second challenge in designing the sample for slum upgrading evaluations is choosing whether and how to stratify the sample frame. One goal of stratified sampling is to insure adequate representation from subpopulations. For instance, evaluators may want to stratify the sample of neighborhoods to ensure regional or district-level representation. This could be relevant if the intervention is suspected to have different features or expected outcomes in different regions of the city or country. Similarly, if the impact evaluation is designed to focus on subgroups such as an ethnic minority, stratified sampling may be the only way to effectively assure that an ethnic minority is adequately represented. In this case, the evaluator must decide whether to ensure an adequate sample of minority households from within neighborhoods (in which case household ethnicity becomes the strata in the second-stage sample) or to seek an adequate sample of neighborhoods with significant minority populations (in which case strata would be defined by the ethnic composition of neighborhoods for first-stage sampling).

In deciding between these approaches, the first consideration is the degree of residential segregation across neighborhoods. Clearly, stratification based on neighborhood racial or ethnic composition is meaningless if neighborhoods are uniformly mixed. Likewise, ensuring representation of ethnic groups within each neighborhood could be impossible if neighborhoods are highly segregated. If residential patterns fall somewhere in between the two extremes, the decision should be based on whether evaluators anticipate greater within- or between-neighborhood ethnic differences in the impact of upgrading. For instance, certain interventions may function differently in different neighborhood types (tenure security could be irrelevant among certain ethnicities with well-functioning systems of informal rights). In other cases, evaluators may be concerned that ethnic minorities are systematically excluded from otherwise well-functioning programs. Another consideration is the degree of residential segregation *within*

neighborhoods. Because slum upgrading is a neighborhood-level intervention and the benefits of certain intervention types will be spatially concentrated, it is arguably optimal to aim for a within-neighborhood sample that is spatially representative. This objective would favor stratifying based on neighborhood rather than household characteristics, especially if there is inadequate information on the exact location of households within neighborhoods.

Stratified sampling can also be employed to increase the precision of estimation when heterogeneous impacts are anticipated across predictable neighborhood characteristics. In this case, the variability within groups should be lower than the variability in the population as a whole, so the sample will have more statistical precision. Relevant strata will generally be specific to the outcomes of interest and intervention location, but slum upgrading evaluations should consider potential neighborhood variation in both the nature of project implementation and anticipated level of impact. Neighborhood characteristics determining anticipated level of impact include direct measures of degree of baseline outcome such as access to water or number of paved roads, as well as indirect indicators of neighborhood benefits such as ethnic composition and poverty.

With respect to project implementation, it may be beneficial to stratify the sample of neighborhoods by size of population since larger communities may have greater difficulty implementing upgrading projects and monitoring quality of supervision or participation by residents or authorities. Similar parameters include population density and age of the neighborhood (average residential tenure or year of community-formation), which could determine both the potential impact and efficacy of implementation. Finally, if the program is characterized by differences in interventions across neighborhoods, evaluators should stratify the neighborhood sample on project type. For instance, if a subset of neighborhoods in a citywide road-paving intervention also receives sewerage upgrading, the sample should be stratified on number of interventions to ensure enough statistical power to assess the separate effect of the two intervention types.

If cross-section survey data is being collected ex-post, it is potentially relevant to stratify the sample of neighborhoods according to the number of years since the upgrading project was completed. In general, evaluations of recently completed projects are likely to find a smaller effect because the full benefits of upgrading have yet to be realized. As a result, it could be useful to ensure an adequate sample of neighborhoods to gauge both short-run and long-run impacts in order to accurately project the stream of future benefits. Meanwhile, a finding that neighborhoods in which the program was recently completed have larger program effects could provide important information about the sustainability of the program.

Minimal required sample sizes will depend on the sample design, the anticipated effect sizes of the focal outcomes and on the methodology being employed. When a simple random sample is collected (for instance, if projects are focused exclusively in one urban area), rough approximations of minimum sample sizes can be calculated with standard power calculations given an anticipated level of impact. For instance, evaluators may determine that the sample should be large enough to detect a proportional change as small as 5% in the focal outcome. In the case of before-and-after comparisons of the project population or ex-post comparisons of treatment and control groups, it would then be necessary to collect a sample of at least 1500 respondents. However, if a minimum difference of 10% is acceptable then the corresponding sample size could be reduced to less than 500.[7] Disaggregating estimates by demographic or regional characteristics requires a corresponding increase in the sample size. Hence, stratified samples designed to capture heterogeneity of impact across subpopulations must take into account the relevant number of divisions in order to calculate a minimum sample size.

4.2 Overlooked areas of potential impact

Clearly, the set of potential direct and indirect outcomes of slum upgrading interventions is large. Appropriate outcome measures within each category will depend on the nature of specific interventions and focal issues in the communities as well as the quality of available data. For guidance on selecting a set of anticipated effects, evaluators should look to previous evaluations in similar settings in addition to theoretical predictions. The following are potential individual and community outcomes that may be relevant for particular slum upgrading interventions, and that have received little attention in the past. We recommend considering these possibilities in addition to standard areas of impact when designing impact evaluations.

Fertility

Upgrading projects may have unintended consequences on family formation through fertility and/or marriage and divorce. This could arise for many reasons depending on the exact nature of the intervention. In the case of land titling programs, there may be important changes in the incentives to bear children or divorce due to shifts in the inter-household allocation of

[7] This approximation is based on using the Z statistic to compute the difference between two population proportions under the following assumptions: (1) sample sizes are equal for the project and control groups; (2) proportions are distributed equally around p = 0.5, which requires the largest sample size; and (3) a two-tailed of the null hypothesis p1 = p2. Under these assumptions, a sample size of 1536 is required to detect a proportional change of 5%. As a rule, it is simpler to base this type of ex-ante power calculation on expected proportional changes as opposed to changes in mean outcomes, which requires information on the standard deviations of the outcomes within the two populations.

resources or ability for parents to secure old-age care through inheritance. For instance, Field (2003) finds evidence of changes in fertility rates following a large land titling program in urban Peru. In the case of health- or education-related slum interventions, changes in infant and child morbidity may alter the value of childbearing.

Information on fertility rates can be gleaned from household rosters, although more complete information can be gathered by collecting birth histories of adult women in the household. In addition, desired fertility may be collected to separate change in birth timing from fertility levels. Marriage and divorce can be similarly measured through standard survey questions. An important caveat is the ability to identify changes in family composition through separation or co-habitation.

Residential segregation

Because slum upgrading intervention may have significant impact on residential mobility, it is important to consider potential changes in neighborhood segregation by race, class or ethnicity. For instance, the community may experience a gradual shift towards wealthier households occupying housing units closer to the water supply, electricity, or roads. On the one hand, some may argue that such trends could have important implications for the program's ability to sustain accessibility and income targeting goals in the long run. On the other hand, depending on the initial assignment of property rights, benefits may flow to the initial owners. The impact on the initial residents of gentrification may be either positive or negative.

Segregation can be tracked by collecting basic characteristics of new occupants (including commercial properties) of housing units that have been abandoned by residents included in the baseline sample, and by following original residents who move. Residential location information from GIS data can help evaluators construct more detailed indicators of neighborhood segregation. Finally, census data may provide decennial information on changes in ethnic and racial composition among small geographic units, if the intervention coincides with these larger data collection efforts.

Formal sector integration

Slum upgrading may have important indirect benefits of facilitating formal sector integration in many areas. Land titling projects provide the most straightforward example. Once residential property rights are secured and land titles registered, it may become easier for home owners to borrow from formal sector lenders, open bank accounts, obtain formal sector employment, participate in government aid programs (pension plans), utilize local health care resources, and enroll their children in school. This integration may also result from reduced fear

or shame of interaction with authorities that motivates community members to participate in the formal sector, or by increasing the penalties associated with informal sector activity.

Evaluations should monitor all of these patterns by asking participants about registration of entrepreneurial activities and property transactions (rental markets), participation in public programs and use of public agencies, and the nature of credit market transactions.

Political enfranchisement

Slum upgrading may affect political enfranchisement including voter participation and local interest in political activism. Some of this could occur rather mechanically if land regularization lowers the cost of voter participation by enabling residents to register at local polls. In some parts of the world, voter registration requires proof of home ownership, limiting the ability of informal urban residents to participate in local elections. Political participation could also be effected by slum upgrading either through psychological channels or through changes in expectations regarding the government's commitment to the neighborhood. For instance, residents are frequently deprived of the status of full citizenship in the eyes of many institutions that require documentation showing proof of a permanent, legal residence, and a demonstrated ability to pay bills.

To monitor improvements in these outcomes, evaluators can generally collect external data on voting rates in local elections, which in most parts of the world is publicly available. Household surveys can collect information on voter participation and other forms of political activity, knowledge of government organizations' roles and responsibilities, and perceptions of the responsibility and performance of these organizations. At the community-level, evaluators could potentially collect information on the existence and nature of local political organizations and community-level activities. Qualitative data or open-ended survey questions may be most useful in capturing changes in feelings of political inclusion or exclusion.

Local governance

Evaluators also want to consider the impact of slum upgrading projects on local governance. Changes in the nature of local governance can be monitored with survey questions on the role of local leadership. To do so, community-level questionnaires may be needed to identify local political bodies, positions, and history. In community-driven development projects, it is particularly relevant to monitor closely the role of local leaders and community governance procedures. These projects could also change the perceptions of local governments about whether they have responsibility for serving these communities.

Intra-household bargaining and gender issues

Land titling programs have the potential to shift within-household distribution of resources, either intentionally or unintentionally. For instance, infrastructure improvements may have gender-specific benefits, for instance reducing the housework burden of women by providing clean water close to the home. Land titling programs may have significant impact on gender relations in the community depending on how property rights are allocated across husbands and wives. To monitor gender-specific outcomes of slum upgrading interventions, evaluators should collect household-level data from both male and female respondents whenever possible. In addition, project participation data should be broken down by gender, including which names appear on property titles, and who pays for community resources, and who participates in managing local resources. In urban communities it is frequently relevant to look separately at the impact of infrastructure improvements on female-headed households.

Mental health, including stress and depression

Evaluations of health impacts of slum upgrading projects frequently center on potential reductions in infectious and parasitic diseases from new sanitation measures, and use of local health clinics. However, significant welfare benefits may operate through non-traditional measures of well-being, such as changes in the psychological burden of stress, hostility and depression. These may be direct or indirect effects of residential improvements. For instance, reductions in crime rates may have a large benefit in terms of reducing stress. Provision of community infrastructure may also have direct psychological benefits of increasing residents' optimism about the future and reducing the stress and hostility associated with social exclusion. Impact evaluations of the Moving to Opportunity program in which residents of inner-city slums in Chicago were relocated to middle-income neighborhoods, found significant psychological benefits of neighborhood upgrading (Kling et al., 2004a, 2004b; Liebman et al., 2004).

Such outcomes can be monitored with relative ease through household surveys that include specially designed mental health modules, generally considered to provide reliable data on psychological well-being. These modules include sets of basic questions about respondent behavior and preferences (feeling sad, crying a lot, not feeling like eating) that allow evaluators to construct common indicators of depression and anxiety. Data from questions about future expectations and optimism can also be used to monitor the psychological impact of particular interventions, in conjunction with self-reported health data and household data on savings and investment.

Informal taxes

In conducting accurate cost recovery analysis, it is important to collect information on the cost of related services before and after the intervention. Cost data collected with this intention is frequently incomplete, since evaluators rarely ask residents about payments made in-kind and through labor exchange. These data should be collected by asking residents about all out of pocket payments for use of community resources, participation in labor exchange and any out-of-pocket payments made to community organizations.

Time use

Because slum upgrading can impact residents' time allocation for many reasons, it may be most efficient to collect time use data from survey respondents. For instance, providing water and sanitation services may significantly reduce the amount of time women in the community spend collecting water or disposing of trash. Similarly, upgrading may be associated with changes in labor supply or neighborhood organizations responsible for building or supplying such services. Finally, new roads could lead to significant reductions in commuting time.

One method of collecting time allocation data is to ask survey respondents about hours spent in specific activities or areas of hypothesized impact, including community projects, household production (cooking, gathering water), labor and leisure. Alternatively, a module can be added to the survey in which respondents are asked to report hourly activities from the day prior to the survey. In both cases, since not all household members can be interviewed, it is typical to ask only the household head, but it may be important to also collect data from the spouse. As in all parts of the survey, it is critical to include an indicator of the particular survey respondent or household member to which the hours correspond.[8] If the latter method is used, it is important to record the day of week

Credit market demand and access

Because slum upgrading can lead to significant increases in the value of housing, these interventions may be associated with significant changes in both residents' demand for credit for housing improvements and banks' willingness to use housing as collateral for loans. Both areas of impact should be monitored by collecting data on household members' demand for and access to both formal and informal credit. For instance, a comprehensive survey questionnaire should ask

[8] Although this may seem obvious, it is a surprisingly common flaw in survey data. Questionnaire-wide rules regarding which household member should answer each section are insufficient to guarantee an accurate mapping from household rosters to individual-level survey data. If the assigned person is unavailable, survey-takers may make unidentifiable substitutions, thereby contaminating all of the data on these modules. To prevent confusion, at the start of each module, a separate question should identify the relationship of the respondent according to his or her person number on the household roster.

the household head (and wife, space permitting) about each credit request made over the past one to three years (depending on lag since baseline survey) and the amount granted. In each case, respondents' should be asked specifically about their use of property titles as collateral in order to track changes in banks' willingness to lend to the poor.

4.3 Evaluation issues in slum upgrading projects

4.3.1 Evaluation issues particular to urban settings

Mobility

One defining feature of urban areas, particularly slums, is high rates of residential mobility. Mobility presents several challenges to project evaluation. For instance, sustainability of infrastructure may become more complicated if individuals engaged in setting up the project leave the community over time. In these settings, an important characteristic of sustainability is whether organizations established to maintain and implement the infrastructure are likely to survive personnel turnover. Mobility also generates high rates of survey attrition, which may bias estimates of program impact. Finally, because upgrading programs affect housing markets, residential mobility may be endogenous to the program, and therefore an important direct outcome of interest.

Since survey attrition can lead to bias, evaluators should take steps to minimize attrition. One simple technique is to tell baseline survey respondents that they will be paid for participating in the follow-up survey. While this may be insufficient to prevent people from moving, evaluators could also encourage people to leave contact information with friends or family or contact survey-takers in the event that they move. A second technique is to lower the cost of follow-up in the second round by collecting detailed information on people in the first round, such as lists of others who would know where they were living as well as contact information for these people, including phone numbers where applicable.

Since mobility is a potentially relevant project outcome, movements in and out of sample households should be measured with care. Follow-up surveys should keep track of the nature and timing of all household composition changes by asking respondents to make a list of all migrants, including the date, destination and reason for migration. In addition, careful attention should be paid to the functioning of land markets as the amount of rental and sales before and afterwards.

Similarly, the survival of maintenance organizations should be measured as a unique project outcome. One implementation question that seemed worth evaluating was maintenance

of latrines. Because slum upgrading projects may experiment with different approaches to maintenance, such as creating private property rights on public goods or tying community-level subsidies to infrastructure maintenance, evaluations should keep careful track of variation in implementation strategies with an eye towards learning about the relative efficacy of specific approaches.

Rural-urban ties

Because of high rural-urban migration rates in many areas of the world, there are frequently strong linkages between rural and urban slum populations. This is relevant for project evaluation for the following reasons. First, urban residents may spend considerable money on remittances to rural areas. Failure to account for these transfers could lead evaluators to underestimate the income effects of the program as well as the overall welfare benefit. For instance, if middle-income program beneficiaries send more transfers to very poor rural areas of the country, the poverty targeting of the program may be higher than a standard impact evaluation implies, though not at the community level.

Linkages between rural and urban areas should be identified at baseline and monitored through follow-up surveys. This includes monetary transfers to relatives in rural areas of the country, and mobility of individuals to and from the area.

Informal sector

In slum areas as in rural villages, most of the population is engaged in informal labor. There are also frequently well-developed informal sector real estate and financial markets. What is unique about many urban settings is that these sectors frequently operate side-by-side with formal sector establishments. While in rural areas of developing countries, close to 100% of credit markets are informal, many urban residents borrow from both formal and informal sources, and business owners may be "partially registered", in terms of having some employees on the formal payroll and others not.

This is important to keep in mind for measuring various aspects of economic well-being and level of market activity. First, informal sector activities may not be included in survey data if survey questions are not carefully designed. Distinguishing between formal and informal sector activity requires even more careful questionnaire design. In addition, as discussed in section 1.4, because an important focus of slum upgrading projects is residential formalization, the degree of formal sector integration may be an important outcome of slum upgrading activities. To the extent that the rate of formal sector participation has externalities and public sector value in

terms of strengthening institutions and government management capabilities, we should view this as an independent welfare benefit of neighborhood upgrading.

Measurement should always follow informal sector activity. Because certain upgrading activities in particular may stimulate transitions from formal to informal sector activity, it is critical to measure substitution of activities across sectors. For instance, an increase in formal borrowing is likely to be accompanied by a reduction in loans from friends and family members. Accurately measuring the extent to which slum upgrading projects reduce credit constraints therefore requires calculating the net effect of these two changes.

Population heterogeneity

Another important feature of urban populations is the potential regional, racial and ethnic population heterogeneity resulting from emigration from various parts of the country and abroad. This feature is relevant for evaluation design for three reasons. First, ethnic tensions could be an important factor in the success of specific projects, and therefore relevant for understanding success and failure rates of specific interventions. This is particularly true in the case of interventions designed to build social cohesion or encourage collective action. The ethnic make-up of a community may also be relevant for understanding patterns of residential mobility and segregation that result from slum improvements. Similarly, it may be ideal to stratify evaluations on race or ethnicity in order to best gauge the impact of the program.

For some variables, it may be important to disaggregate impact by gender, race or ethnicity. For instance, community participation and usage rates may become concentrated among a certain ethnic group. In addition, evaluations should keep track of changes in community composition by race and ethnicity, and, when possible, residential segregation.

4.3.2 Evaluation issues particular to public goods

Spillovers

Particularly because population density is high in urban areas, spillovers are an important evaluation issue. Examples of positive spillover effects on neighboring communities in the context of slum upgrading projects include: higher property values, crime reduction, and access to public goods such as roads and parks that may be used heavily by members of neighboring communities. Although externalities are generally presumed to be positive, possible negative spillovers include: diversion of criminal activities or waste disposal to neighboring areas, water runoff or environmental waste from construction.

While the presence of large spillovers can dramatically raise the value of a slum upgrading project, this feature also confounds standard evaluation methodologies. In these cases, careful approaches to impact assessment are necessary. First and foremost, selecting a control group is more difficult when externalities are large. When program benefits extend to people outside of the assigned treatment group, standard randomization procedures face two potential problems. First, people in the control group may be among those who benefit from the program. Hence, differences in outcomes before and after the intervention do not necessarily reflect confounding time trends or mean reversion as in classic experimental designs, but may be a direct result of the intervention. In this case, the DID estimator is no longer a valid identification strategy. Second, even if the control group is isolated from the treatment area, a standard DID estimator will fail to account for the added benefit of the program to non-participants, and thereby underestimate the total value of the intervention. In many cases, the value of spillovers far outweigh the value of the program to direct beneficiaries, hence capturing the total impact is potentially important.

The magnitude of spillovers can frequently be captured and measured if evaluation methodologies are designed with this in mind. Kremer and Miguel (2004) discuss approaches to identifying program impact in the presence of spillovers. There are two basic approaches. First, in all types of evaluation, basic comparison groups should be chosen with enough social and physical distance so as to prevent cross-unit contamination via spillovers from treatment to control groups. While there are no clear boundaries, the disadvantages of choosing geographically distant areas must be carefully weighed against the benefit of minimizing contamination of the control group. In all cases, evaluators must spend considerable time verifying with background data that control areas are similar in every observable dimension that is available, unless there is a clear argument of orthogonality. Second, if possible, spillovers can be directly measured by randomly assigning treatment densities by randomizing across units as well as within units (Miguel and Kremer, 2004; Duflo and Saez, 2003).

When cross-unit randomization of treatment density is not possible, it may be possible to estimate the magnitude of externalities by identifying individual variation in predicted level of externality. One such variable used by Kremer and Miguel (2004) is the geographic distance of non-treated households from treated households. For community level improvements, this method is only useful if there is significant variation within neighborhoods in terms of access to and direct benefits of infrastructure improvements. For instance, to monitor neighborhood externalities from a latrine intervention in terms of the spread of parasitic disease, untreated households within the neighborhood could be characterized by the fraction of neighbors within a two-block radius that received latrines. See Miguel and Kremer for a discussion of this approach.

In the case of community-wide treatment (for instance, roads placed throughout the community), neighboring communities need to be monitored in order to gauge the degree of externalities.

When randomization is not possible, a simpler strategy is to choose two levels of control groups, one that is within the relevant project area or slum, and a second that lies outside of the slum that is being upgraded. The first control group would differ from the treatment group only in so far as they are expected to benefit from the intervention. For instance, households inside of treatment areas for sewage connections with access to a water connection can be compared to those households that lacked indoor water supply before the program. The second control group would then consist of households in a neighborhood far enough away from the intervention that externalities are not a concern. In this manner, a difference-in-difference-in-difference (DIDID) estimate can be used to eliminate some of the bias that may arise from choosing a control group that is very different from the treatment group.

Contamination

Accurately calculating program impacts requires that the control group not alter its behavior in anticipation of future upgrading initiatives in their own neighborhoods. Because urban areas are characterized by high population density and connectivity to sources of local information, this type of contamination is frequently a concern for project evaluation in urban settings.

When contamination and anticipatory effects are hard to avoid, randomized evaluations may be cleaner in a pilot version of the program conducted among a small set of households in remote areas. At this stage, it may be possible to minimize knowledge of the program, and to select control communities far enough away to avoid contamination. However, it is again important to keep in mind the tradeoff between selecting a community that is similar to the intervention area versus selecting a community in which there are no anticipatory effects. In many cases – for instance, nationwide programs – the latter will be impossible to accomplish since most communities will be aware of the program. In these cases, it is recommended that evaluators monitor households' expectations with carefully designed survey questions.

4.3.3 Evaluation issues particular to slum upgrading project implementation

Crime

In many parts of the world, urban slums are characterized by high rates of crime and black-market activity, and local activities may be managed by criminal gangs or leaders. This is relevant for project evaluation since it may be difficult to collect accurate information on income

and employment. In addition, criminal organizations may present significant barriers to project implementation, so may be a relevant neighborhood feature in assessing program impact and extrapolating evaluation results to other settings.

Where misreporting is likely due to high rates of criminal activity, socioeconomic outcomes should be measured through other survey questions such as expenditures and assets. Since crime itself is a potentially important outcome, baseline and follow-up surveys should collect detailed information on the nature of local crime and violence. Here, local authorities may be useful in providing statistics on police activity and crime rates over time. More likely, much of this activity will not be reported to local authorities so must be collected in household surveys that ask project participants about perceptions of security and about incidence of crime victimization. To do so, it is advisable to include questions on general feelings of security in addition to reports of specific and relevant criminal acts. For instance, a questionnaire could include questions such as whether the respondent would feel comfortable walking through the neighborhood alone at night, whether the respondent has been mugged during the last month, and whether any goods have been stolen from inside the property over the past month.

There may be important gender differences with respect to feelings of safety, so questionnaires may be more accurate if they target male and female respondents separately. Similarly, the level of crime in neighboring communities may be an important outcome to monitor through survey questions.

Multiple simultaneous interventions

One of the principal difficulties in conducting impact evaluations of slum upgrading projects stems from the fact that projects typically comprise several simultaneous interventions. Although it is always possible to focus exclusively on the net benefit of a set of interventions, the goal of impact assessment is not only to identify the total project effect, but also to evaluate the relative value of each project component and the value of interacting specific features. This is relevant both for scaling up pilot studies and for predicting the benefits for similar projects in other areas. Although we are interested in the overall impact of slum upgrading, there is still much to be learned about the optimal design of these programs. Since the programs are typically complex and have many different components, it is valuable to determine which components are the real drivers of change not only in terms of independent benefits, but also in terms of complementarities between specific design features.

For projects that involve multiple interventions one would ideally use randomization procedures to assess both an overall package of interventions and a number of variants that

include some but not all components in order to assess the roles of different components. In particular, neighborhoods could be randomly assigned different slum upgrading packages that include a set of varying project components. If there are sufficient neighborhoods across which the intervention will take place and it is feasible to vary the projects widely, both the separate effects and interaction effects could in theory be identified. Obviously, the fewer project components the more feasible is such a strategy.

Another identification strategy involves phasing in different project components. In some instances, phase-in will be a natural feature of project implementation since it is generally infeasible to activate treatments simultaneously. In these cases, project components may be phased into neighborhoods with varying schedules somewhat arbitrarily, such that at one point in time different communities have different sets of treatments. In other cases, evaluators may be able to convince program administrators to vary either the speed or order in which particular features are added to the neighborhood upgrading. For instance, if a slum upgrading project involves painting houses and paving roads over the course of a year, it may be easy to convince project administrators to begin with roads in a randomly chosen set of communities and paint in the remainder. Then in the second stage of the project the second intervention would complete the upgrading in all neighborhoods. In this manner, the relative if not absolute impact of paved roads versus painted homes could be estimated using the DID estimator.

When evaluators do not have sufficient control to implement randomized or phased-in project components, they can attempt to assess the relative importance of various project features by identifying neighborhood characteristics related to the relative value of specific projects. The prior availability of certain program features is the most straight-forward approach. For instance, if a slum upgrading project is providing water, sewage and roads to a number of urban slums some of which already have some of this infrastructure, the relative impact of each can, with some assumptions, be identified. A more subtle evaluation design involves identifying community characteristics related to the relative benefit of roads versus water. For example, communities on the mountainous side of the city may have plentiful access to fresh water relative to communities on the coastal side, but greater need for roads due to rough terrain.

5 Conclusions

We argued that while all programs should be subject to process evaluations, not all programs should be subject to impact evaluations. In some situations the assumptions underlying quasi-experimental methodologies will be plausibly satisfied, implying quasi-experimental impact evaluations can naturally be applied. In addition, for

a subset of policy relevant questions that have been identified as priorities by the Bank, undertaking randomized evaluations would be very useful. Given that evaluation is a global public good and that there is a significant cost to doing randomized evaluations for specific task managers and researchers, it therefore makes sense for the Bank to centrally ensure that some percent of activity is randomized (this percentage figure is now very close to zero).

With regard to slum upgrading evaluations in particular, careful consideration should be given to the wide range of individual- and community-level outcomes that could be monitored, especially as there exist several important questions which previous studies of slum upgrading projects have not been able to address (such as sustainability concerns). Issues specific to urban settings, public goods, and implementation should be taken into account when planning rigorous evaluations of slum upgrading projects.

Bibliography

EVALUATIONS AND CASE STUDIES

Abelson, Peter. 1996. "Evaluation of slum improvements: Case study of Visakhapatnam, India." Cities 13(2): 97-108.

Abramson, Daniel Benjamin. 1997. "Marketization and Institutions in Chinese Inner-city Redevelopment: a Commentary of Lu Junhua's Beijing's Old and Dilapidated Housing Renewal." Cities 14 (2): pp 71-75.

Angel, Shlomo with Thipparat Chirathamkijkul. 1983. "Slum Reconstruction: Land Sharing as an Alternative to Eviction in Bangkok." In Land for Housing the Poor. Edited by Shlomo Angel, et al, 461-472. Bangkok: Select Books.

Bamberger, Michael, Gonzalez, E. and U. Sae-Hau. 1982. "Evaluation of Sites and Services Projects: the Evidence from El Salvador." Washington, DC: World Bank.

Bamberger, Michael and Eleanor Hewitt. 1986. "Monitoring and Evaluating Urban Development Programs: A Handbook for Program Managers and Researchers." World Bank Technical Paper No. 53. Washington, D.C.

Bamberger, Michael, Sanyal, Bish and Nelson Valverde. 1982. "Evaluation of Sites and Services Projects: the Experience from Lusaka, Zambia." Washington, DC: World Bank.

Bijlani, H.U. 1988. "Strategies for urban shelter : The improvement of slums, squatter settlements and sites-and-services." Habitat International, 12(4): 45-53.

Coit, Katharine. 1998. "Housing policy and slum upgrading in Ho Chi Minh City." Habitat International 22(3): 273-280.

Dundar, Ozlem. 2001. "Models of urban transformation: Informal housing in Ankara." Cities, 18(6): 391-401.

Ghafur, Shayer. 2001. "Beyond homemaking: the role of slum improvement in homebased income generation in Bangladesh." Third World Planning Review 23(2): 111-135.

Imparato, Ivo and Jeff Ruster. 2003. "Slum upgrading and participation: Lessons from Latin America." Directions in Development Series. Washington, D.C.: World Bank.

Kalim, Syed Iqbal. 1983. "Incorporating Slum Dwellers into Redevelopment Schemes: The Jacobs Lines Project in Karachi." Land for Housing the Poor. Edited by Shlomo Angel, et al., 461-472. Bangkok: Select Books.

Kapoor, Mudit, Lall, Somik V., Lundberg, Matthias K.A., and Shalizi, Zmarak. May 2004. "Location and Welfare in Cities: Impact of Policy Interventions on the Urban Poor." World Bank Policy Research Working Paper No. 3318. Washington, D.C.

Keare, Douglas H. and Parris, Scott. 1982. "Evaluation of Shelter Programs for the Urban Poor: Principal Findings." World Bank Staff Working Papers No. 547. Washington, D.C.

Lu, Junhua. 1997. "Beijing's Old and Dilapidated Housing Renewal." Cities 14 (2): 59-69.

Miah, Md. Abdul Quader and Karl E. Weber. 1990. "Feasible slum upgrading for Dhaka." Habitat International 14(1): 145-160.

Mukhija, Vinit. 2001. "Upgrading Housing Settlements in Developing Countries: The Impact of Existing Physical Conditions." Cities 18(4): 213-222.

Munarriz, Mayu T. 1988. "Socioeconomic Changes in Tondo Foreshore: An Evaluation of a Slum Upgrading Project." Philippine Review of Economics and Business 25(3-4): 255-70.

Schoorl, J. W., J. Van der Linden, and K. S. Yap. 1983. Between Basti Dwellers and Bureaucrats: Lessons in Squatter Settlement Upgrading in Karachi. New York: Pergamon Press.

Serag-el-Din, Hany B. 1990. "The Effects of Combined Upgrading and New Development Schemes on Housing Patterns of the Site: Case Study of Ismailia Project in Egypt." Housing Science 14 (4): 259-272.

Verma, Gita Dewan. 2000. "Indore's Habitat Improvement Project: Success or failure?" Habitat International 24(1): 91-117

Werlin, Herbert. 1999. "The Slum Upgrading Myth." Urban Studies 36(9): 1523-1534.

World Bank, OED, 1996. "Kenya - Development of Housing, Water Supply and Sanitation in Nairobi." Impact Evaluation Report No.15586. Washington, D.C.

World Bank, Operations Evaluation Department, 1995. "Indonesia - Enhancing the Quality of Life in Urban Indonesia: The Legacy of Kampung Improvement Program." 1995. Impact Evaluation Report No. 14747. Washington D.C.

Yun, L.W., Yusof, K, 1991. "Services for urban poor families in Kuala Lumpur, Malaysia: A case study." Child Welfare 70(2): 293-292.

METHODOLOGY

Kaufman, Daniel and John M. Quigley. 1987. "The consumption benefits of investment in infrastructure: The evaluation of sites-and-services programs in underdeveloped countries." Journal of Development Economics 25(2): 263-284.

Keare, Douglas H. and Parris, Scott. 1982. "Evaluation of Shelter Programs for the Urban Poor: Principal Findings." World Bank Staff Working Papers No. 547. Washington, D.C.

Krige, Skip, Schur, Michael, and Gerd Sippel. 1998. "The identification of towns in the Free State for an urban upgrading and development programme: a proposed method for consideration." Development Southern Africa 15(3): 361-377.

Kumar, A. 1991. "Delivery and management of basic services to the urban poor: the role of the urban basic services, Delhi." Community Development Journal 26(1): 50–60.

Mukhija, Vinit. 2001. "Enabling Slum Redevelopment in Mumbai: Policy Paradox in Practice," Housing Studies 16(6): 791-806.

Salmen, Lawrence F. 1987. Listen to the people: Participant-observer evaluation of development projects. New York: Oxford University Press/World Bank.

HOUSING PRICES

Crane, Randall, Daniere, Amrita, and Stacy Harwood. 1997. "The Contribution of Environmental Amenities to Low-Income Housing: A Comparative Study of Bangkok and Jakarta." Urban Studies 34(9): 1495-1512.

LAND TENURE AND TENURE SECURITY

Bassett, Ellen M. 2004. "Tinkering with tenure: the community land trust experiment in Voi, Kenya," Habitat International. In press. Available online 27 February 2004.

Field, Erica (2003). "Entitled to Work: Urban Tenure Security and Labor Supply in Peru." Princeton University Research Program in Development Studies Working Paper #220, November 2003.

Field, Erica (2003) "Fertility Responses to Urban Land Titling Programs: The Roles of Ownership Security and the Distribution of Household Assets." Manuscript, Harvard University, October 2003.

Field, Erica (2003) "Property Rights, Community Public Goods and Household Time Allocation in Urban Squatter Communities" *William and Mary Law Review,* February 2004, 45(3): 837-887.

Field, Erica (2004) "Property Rights and Investment in Urban Slums." Forthcoming, *Journal of the European Economic Association Papers and Proceedings.*

Field, Erica and Maximo Torero (2003). "Do Property Titles Increase Credit Access among the Urban Poor? Evidence from Peru" Manuscript, Harvard University, January 2004.

Jimenez, Emmanuel. 1983. "The Magnitude and Determinants of Home Improvements in Self-Help Housing Manila's Tondo Project." Land Economics 59: 70-83.

Jimenez, Emmanuel. 1988. "Urban services and rural infrastructure." Finance and Development 25(3): 6-8.

Keare, Douglas and Emmanuel Jimenez. 1983. "Progressive Development and Affordability in the Design of Urban Shelter Projects." World Bank Staff Working Papers No. 560. Washington D.C.

Hoy, Michael and Emmanuel Jimenez. 1996. "The impact on the urban environment of incomplete property rights." World Bank Poverty, Environment and Growth Working Paper No.14. Washington D.C.

Lanjouw, J. O. and Philip Levy (2002) "Untitled: A Study of Formal and Informal Property Rights in Urban Ecuador," *The Economic Journal.* Vol. 112, pp. 986-1019.

ADMINISTRATION

Lee, Yok-Shiu F. 1998. "Intermediary institutions, community organizations, and urban environmental management: The case of three Bangkok slums." World Development 26(6): 993-1011.

Olken, Benjamin (2005). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." Manuscript, Harvard University, February 2005.

Otiso, Kefa M. 2003. "State, voluntary and private sector partnerships for slum upgrading and basic service delivery in Nairobi City, Kenya." Cities 20(4): 221-229.

WATER AND ENVIRONMENT

Enabor, Bolu, Sridhar, M.K.C., and I.O. Olaseha. 1998. "Integrated Water Management by Urban Poor Women: A Nigerian Slum Experience." International Journal of Water Resources Development 14(4): 505-512.

Hasan, A. 1993. "Karachi's poor neighborhoods achieve low-cost sanitation." Water Resources Journal (September): 82–83.

Hirotsugu Aiga and Takusei Umenai. 2002. "Impact of improvement of water supply on household economy in a squatter area of Manila." Social Science & Medicine 55(4): 627-641.

McPhail, A.A. 1993. "The 'five percent rule' for improved water service: can households afford more?" World Development 21(6): 963–973.

Perla, Martha. 1997. "Community composting in developing countries." BioCycle 38(6): 48-51.

HEALTH AND EDUCATION

Awasthi, Shally, Nichter, Mark, and V.K. Pande. 2000. "Developing an Interactive STD-Prevention Program for Youth: Lessons from a North Indian Slum," Studies in Family Planning 31(2): 138-50.

Harpham, Trudy and Carolyn Stevens. 1992. "Policy directions in urban health in developing countries—The slum improvement approach," Social Science and Medicine 35(2):111-120.

Khullar, Mala and Shayam Menon. 1996. "Innovative Approaches in Early Childhood Education: Evaluation Report." Aga Khan Foundation, ED414048, Geneva.

Kremer, Michael and Edward Miguel (2004). "Worms: Identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72(1), 159-217.

Stanton, Bonita F., Clemens, John D., Khair, Tajkera, Khatun, Khodeza, and Dilwara Akhter Jahan. 1987. "An educational intervention for altering water-sanitation behaviours to reduce childhood diarrhoea in urban Bangladesh: Formulation, preparation and delivery of educational intervention." Social Science and Medicine 24(3): 275-283.

WOMEN

Baruah, Bipasha, 2004. "Under One Roof." Women & Environments International Magazine Spring/Summer (62/63): 5-7.

Shami, Seteney. 1996. "Gender, domestic space, and urban upgrading: a case study from Amman." Gender and Development 4(1): 17-23.

OTHER

Claudio Acioly with Paul Procee and David Edelman (1999). "Sustainable Urban Development and the Urban Poor in Rio de Janeiro", in 'The Challenge of Environmental Management in Urban Areas', M. Mattingly, E. Fernandes, J. Davila and A. Atkinson (eds), Ashgate, London, UK, pp. 127-138, 1999.

José Brakarz with Margarita Greene and Eduardo Rojas, (2000). "Cities for All. Recent Experiences with Neighbourhood Upgrading Programs", Inter-American Development Bank, 2000.

Alain Durand-Lasserve, Edésio Fernandes, Geoffrey Payne and Marim Smolka, (undated). "Land Tenure for the Urban Poor", in CIVIS, Learning from Cities, Cities Alliance, USA.

Alain Durand-Lasserve and Valérie Clerk (1996). "Regularisation and Integration of Irregular Settlements. Lessons from Experience", UMP Working Paper 6, Urban Management Programme, USA, March 1996.

Duflo, Esther and Emmanuel Saez (2003). "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment" *Quarterly Journal of Economics*, 118, 2003, 815-842.

Jeffrey B. Liebman, Lawrence F. Katz, and Jeffrey R. Kling (2004). "Beyond Treatment Effects: Estimating the Relationship Between Neighborhood Poverty and Individual Outcomes in the MTO Experiment." Manuscript, Harvard University.

Jeffrey R. Kling, Jeffrey B. Liebman, Lawrence F. Katz, and, Lisa Sanbonmatsu (2004a). "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult

Economic Self-Sufficiency and Health from a Randomized Voucher Experiment." Manuscript, Princeton University.

Jeffrey R. Kling, Jeffrey B. Liebman, and Lawrence F. Katz (2004b). "Bullets Don't Got No Name: Consequences of Fear in the Ghetto." In T.S. Weisner, ed., *Discovering Successful Pathways in Children's Development: New Methods in the Study of Childhood and Family Life*, U. of Chicago, 2004.

RANDOMIZED EVALUATIONS

Angrist, Joshua and Alan Krueger (1999) "Empirical Strategies in Labor Economics," in Handbook of Labor Economics, Vol. 3A. Orley Ashenfelter and David Card (Eds.). Amsterdam: North Holland, 1277-1366.

Angrist, Joshua and Alan Krueger (2001) "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," Journal of Economic Perspectives, 15(4), 69-85.

Angrist, Joshua and Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," Quarterly Journal of Economics, 114(2), 533-575.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek (2000) "A Review of Estimates of Schooling/Earnings Relationship, with Tests for Publication Bias," NBER working paper #7457.

Attanasio, Orazio, Costas Meghir and Ana Santiago (2001) "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," mimeo, Inter-American Development Bank.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden (2003) "Improving the Quality of Education in India: Evidence from Three Randomized Experiments," mimeo, MIT.

Behrman, Jere, Piyali Sengupta and Petra Todd (2002) "Progressing through PROGESA: An Impact Assessment of a School Subsidy Experiment in Mexico," mimeo, University of Pennsylvania.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan (2004) "How Much Should We Trust Difference in Differences Estimates?" Quarterly Journal of Economics, 119(1): 249-275.

Besley, Timothy and Anne Case (2000) "Unnatural Experiments? Estimating the Incidence of Endogenous Policies,' Economic Journal, 110(467), F672-F694.

Buddlemeyer, Hielke and Emmanuel Skofias (2003) "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA," Institute for Study of Labor, Discussion Paper No. 827.

Campbell, Donald T. (1969) "Reforms as Experiments," American Psychologist, 24, 407-429.

Card, David (1999) "The Causal Effect of Education on Earnings," in Handbook of Labor Economics, Vol. 3A. Orley Ashenfelter and David Card (Eds.). Amsterdam: North Holland, pp. 1801-63.

DeLong, J. Bradford and Kevin Lang (1992) "Are All Economic Hypotheses False? " Journal of Political Economy, 100(6) (December), 1257-72.

Duflo, Esther (2001) "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," American Economic Review, 91(4): 795-814.

Duflo, Esther and Michael Kremer (2005) "Use of Randomization in the Evaluation of Development Effectiveness," in George Pitman, Osvaldo Feinstein, and Gregory Ingram (editors), Evaluating Development Effectiveness, New Brunswick, NJ: Transaction Publishers.

Duflo, Esther and Emmanuel Saez (2003) "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," Quarterly Journal of Economics, 118: 815-842.

Glazerman, Steven, Dan Levy, and David Meyers (2002) "Nonexperimental Replications of Social Experiments: A Systematic Review." Mathematica Policy Research, Inc. Interim Report/Discussion Paper.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz (2004) "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," Journal of Development Economics, 74(1): 251-268.

Imbens, Guido and Joshua Angrist (1994) "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62(2), 467-475.

LaLonde, Robert (1986) "Evaluating the Econometric Evaluations of Training with Experimental Data," The American Economic Review, 76(4), 604-620.

Meyer, Bruce D. (1995) "Natural and quasi-experiments in economics," Journal of Business and Economic Statistics, 13(2), 151-161.

Miguel, Edward and Michael Kremer (2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," Econometrica, 72(1): 159-217.

Miguel, Edward and Michael Kremer (2003) "Social Networks and Learning About Health in Kenya," mimeo, Harvard University.

Morduch, Jonathan (1998) "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh," mimeo, Princeton University.

Pitt, Mark and Shahidur Khandker (1998) "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" Journal of Political Economy, 106(5), pp. 958-996.

Pritchett, Lant (2003) "It Pays to be Ignorant," forthcoming, Journal of Policy Reform.

Vermeersch, Christel and Michael Kremer (2005) "School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Experiment," mimeo, Harvard University.

Appendix: Additional discussion of randomized evaluations.

In this Appendix[9], we include some additional discussion drawing on the experience of the authors on some lessons which can be drawn from previous randomized evaluations, with the intention of offering guidance for individuals interested in implementing randomized evaluations.

Randomized evaluations are often feasible.

For a variety of reasons, randomized evaluations are more feasible than many people may think. A common perception is that political economy concerns over randomized evaluations may sometimes make it difficult to not implement a program in the entire population. However, these concerns can be, and have in the past been, successfully tackled at several levels. For example, in some cases it might be possible to work with NGOs to do "beta versions" of programs before a program is implemented on a large scale. For larger scale projects, financial constraints often necessitate phasing-in programs over time, and randomization may actually be the fairest way of determining the order of phase-in.

Another potential means of addressing political economy concerns is to acknowledge that in some randomized program evaluations it may be more feasible to randomize among a subset of neighborhoods rather than to randomize among the entire population of neighborhoods. For instance, administrators could select a large group of neighborhoods among which they are relatively indifferent about order, and then randomize within that group. In other words, if, for priority or political reasons, there are certain neighborhoods which cannot be randomized, that is not necessarily inconsistent with doing a randomized evaluation on the sample of other neighborhoods.

The historical record evidences that randomized evaluations can be successfully undertaken by country governments – such as with the Mexican PROGRESA program, the Moving to Opportunity program in the US, and a Honduras land titling program. For the case of the Honduras land titling program, the government of Honduras is currently undertaking a nation-wide land titling program in both urban and rural areas of the country. All neighborhoods in the country will be titling within 18 months, and roughly 15% of neighborhoods were selected for early intervention due to political demands and pilot testing. To facilitate the impact evaluation, being conducted by Erica

[9] This section draws heavily on Duflo and Kremer (2005).

Field and Maximo Torero, program administrators agreed to randomize the order (date of entry) of the remaining 85% of neighborhoods that will be reached by the titling program.

In the case of the Moving to Opportunity (MTO) program, the US Department of Housing and Urban Development (HUD) undertook a randomized trial as part of a national demonstration program. Families in five cities (Baltimore, Boston, Chicago, Los Angeles, and New York) were eligible for participation in the demonstration if they had children and resided in public housing or project-based Section 8 assisted housing in census tracts with a 1990 poverty rate of 40 percent or higher. Families were randomly assigned assistance in moving to wealthier neighborhoods, thus providing an opportunity both to assess the effectiveness of using housing mobility programs to move families to these neighborhoods and to measure the causal impacts of neighborhood attributes on family and youth outcomes for poor families.

However, as we will discuss, below governments are far from being the only outlets through which randomized evaluations can be conducted. NGOs have the potential to play a very valuable role in conducting randomized evaluations.

A more general point is that given a question which has been identified to be of important policy interest, it is likely to be possible for the World Bank to find a context in which a randomized evaluation of that policy or program can be feasibly implemented -- even though for any given World Bank project in a particular setting it may not be feasible to conduct a randomized evaluation. A wide variety of examples provide evidence that randomized evaluations are often feasible and have been conducted successfully.

NGOs are well-suited to conduct randomized evaluations, but will require technical assistance (for example, from academics) and outside financing.

Governments are far from being the only possible outlets through which randomized evaluations can be organized. Evaluation of NGO projects may be a particularly promising way of ascertaining what works. Evidence from the evaluation of NGO programs can then, in turn, provide credible guidance to country governments and organizations such as the World Bank in considering which programs are most effective and should be scaled-up to larger populations.

NGO programs are strong candidates for evaluations for several reasons. Unlike governments, NGOs are not expected to serve entire populations and hence may face

fewer political economy constraints in implementing randomized trials or randomized phase-ins. Financial and administrative constraints often lead NGOs to phase in programs over time, and randomization will often be the fairest way of determining the order of phase-in. Even small NGOs can substantially affect budgets in developing countries. Given that many NGOs exist and that they frequently seek out new projects, it is often relatively straightforward to find NGOs willing to conduct randomized evaluations: hitches are more often logistical than philosophical. For example, a set of recent studies conducted in Kenya have been carried out through a collaboration with the Kenyan NGO Internationaal Christelijk Steunfonds (ICS) Africa: ICS was keenly interested in using randomized evaluations to see the impact its programs are having, as well in sharing credible evaluation results with other stake holders and policy makers.

A second example is a collaboration between the Indian NGO Pratham and MIT researchers which led to the evaluations of remedial education and computer-assisted learning programs (Banerjee et al., 2003). This collaboration was initiated when Pratham was looking for partnership to evaluate their programs; Pratham understood the value of randomization and was able to convey the importance of such evaluations to the schoolteachers involved in the project. However, while NGOs are well placed to conduct randomized evaluations, it is less reasonable to expect them to finance these evaluations. The evaluations of the ICS de-worming programs were made possible by financial support from the World Bank, the Partnership for Child Development, and the US National Institutes of Health (NIH), and the MacArthur Foundation. In the case of the Indian educational programs, Pratham was able to find a corporate sponsor: India's second-largest bank, ICICI Bank, was keenly interested in evaluating the impact of the program and helped to finance part of the evaluation. In general, given that accurate estimates of program effects are international public goods, we argue that randomized evaluations should be financed internationally. International organizations could finance randomized evaluations organized in collaboration between researchers (from those organizations as well as from academia) and NGOs, and the potential is enormous.

Costs can be reduced and comparability enhanced by conducting a series of evaluations in the same area.

Once staffs are trained, they can work on multiple projects. Since data collection is the most costly element of these evaluations, crosscutting the sample can also dramatically reduce costs. This tactic must take into account potential interactions

between programs (which can be estimated if the sample is large enough), and may not be appropriate if one program makes the schools atypical.
National evaluation agencies, such as are starting to emerge in Latin America, could take advantage of similar benefits from trained staff working on multiple projects.

Randomized evaluations can be useful in generating credible cost-effectiveness estimates.

Many of the studies of (for example) educational interventions are limited in that a central policy concern for developing countries is the relative cost-effectiveness of various interventions to increase school participation. Evaluations of cost-effectiveness require knowledge of a program's costs as well as its impact, and comparability across studies requires some common environment. Comparing the impact of PROGRESA's cash transfers and school meals in Kenya is difficult, since it is unclear whether the resulting differences are associated with the type of program or the larger environment. Policymakers are usually left with an unappealing choice between retrospective studies, which allow comparisons of different factors affecting school participation but may yield biased estimates, and randomized evaluations, which yield credible estimates but only for a single programs.

One exception to our general inability to compare cost-effectiveness of credible estimates is the recent set of studies conducted in Kenya. Because the Kenyan programs discussed in this section were conducted in similar environments, cost-effectiveness estimates from these randomized evaluations can be readily compared. Deworming was found to be extraordinarily cost-effective at only $3.50 per additional year of schooling (Miguel and Kremer, 2004). In contrast, even under optimistic assumptions, provision of free uniforms would cost $99 per additional year of school participation induced (Kremer et al., 2002). The school meals program, which targeted preschoolers rather than primary school age children, cost $36 per additional year of schooling induced (Vermeersch and Kremer, 2004).

Randomized evaluations can be designed to test for various channels of influence.

If researchers want to disentangle the channels through which program impacts occurred, one option is to design several treatments in order to break apart the various channels. This may be of particular interest if there are concerns over "promotional" effects.

One example is the ongoing evaluation of a fertilizer intervention by Esther Duflo, Michael Kremer, and Jonathan Robinson. A program in which farmers were offered the opportunity to commit to save had substantial effects on take up of fertilizer. However, the authors were concerned that the program might simply be working by convincing the farmers that an NGO that is active in their area wants them to use fertilizer. This may lead them to either start using fertilizer or to start pretending that they are using it. To test this hypothesis, the authors visited a randomly selected group of farmers, and delivered a speech "endorsing" fertilizer use. By comparing fertilizer take-up rates in the NGO-program treatment group and this "endorsement" group, the authors were able to isolate the direct effect of the NGO program from any potential reputation effect of the NGO itself.

A second example is the evaluation of a savings program ("SEED") by Nava Ashraf, Dean Karlan, and Wesley Yin. In this case the authors were concern that individuals may have saved more not because of the special SEED savings program but just because they were involved with a savings program. Their solution in this case was to create an additional treatment group to which they marketed a "regular" savings program. That is, of the half of these individuals not assigned to the "SEED commitment treatment" group, one-fourth were assigned to the "control" group while one-fourth were assigned a third group-the "marketing treatment" group. Clients in this group were given virtually the same marketing campaign as received by clients in the "SEED commitment treatment" group, except that the marketing was strictly limited to conventional and existing savings products of the participating micro-finance institution. By comparing savings levels of clients in the "SEED commitment treatment" and "marketing treatment" groups, the authors were able to isolate the direct effect of the SEED product from the effect of the marketing campaign.

Randomized evaluations have a number of limitations, but many of these limitations also apply to other techniques.

Many of the limitations of randomized evaluations also apply to other techniques. We here review four issues which affect both randomized and non-randomized evaluations (sample selection bias, attrition bias, behavioral responses, and spillover effects), and we will argue that randomized methods often allow for easier correction for these limitations than do non-randomized methods.

Sample selection problems could arise if factors other than random assignment influence program allocation. For example, people may move from a place without the program to a place with the program. Conversely, individuals allocated to a treatment group may not receive the treatment (for example, because they decide not to take up the program). Even if randomized methods have been employed and the intended allocation of the program was random, the actual allocation may not be. This problem can be addressed through "intention to treat (ITT)" methods by using random assignment as an instrumental variable for actual assignment. Although the initial assignment does not guarantee in this case that someone is actually either in the program or in the comparison group, in most cases it is at least more likely that someone is in the program group if he or she was initially allocated to it. The researcher can thus compare outcomes in the initially assigned group and scale up the difference by dividing it by the difference in the probability of receiving the treatment in those two groups to obtain the local average treatment effect estimate (Imbens and Angrist, 1994). Methods such as ITT estimates allow selection problems to be addressed fairly easily in the context of randomized evaluations, but it is often much more difficult to make these corrections in the case of a retrospective analysis.

A second issue affecting both randomized and non-randomized evaluations is differential attrition in the treatment and the comparison groups: those who participate in the program may be less likely to move or otherwise drop out of the sample that those who do not. Statistical techniques can be used to bound the potential bias, but the ideal is to try to limit attrition as much as possible. All this requires knowing who was present initially, which is much easier with randomized evaluations.

Finally, programs may create spillover effects on people who have themselves not been treated. These spillovers may be physical, as found for the Kenyan de-worming program by Miguel and Kremer (2004) when de-worming interferes with disease transmission and thus makes children in treatment schools (and in schools near treatment schools) less likely to have worms even if they were not themselves given the medicine. Such spillovers might also operate through prices, as found by Vermeersch and Kremer (2005) when the provision of school meals leads competing local schools to reduce school fees. Finally, there might also be learning and imitation effects (Duflo and Saez, 2003; Miguel and Kremer, 2003). If such spillovers are global (for example, due to changes in world prices), identification of total program impacts will be problematic with any methodology. However, if such spillovers are local then randomization at the level

of groups can allow estimation of the total program effect within groups and can generate sufficient variation in local treatment density to measure spillovers across groups. For example, the solution in the case of the de-worming study was to choose the school (rather than the pupils within a school) as the unit of randomization (Miguel and Kremer, 2004), (of course, this requires a larger sample size).

If the unit of observation is above the individual level, concerns about potential common shocks must be taken into account. Common shocks are addressed econometrically through clustering techniques which correct standard errors appropriately; when conducting prospective evaluations, clustering concerns need to be taken into account when considering power calculations and what sample sizes are needed.

One issue that may not be as easily dealt with is that the provision of inputs might temporarily increase morale among students and teachers, which could improve performance. While this would bias randomized evaluations, it would also bias fixed-effect or difference-in-difference estimates. However, it is unclear how serious of an issue this is in practice, whereas we know selection is a serious concern.

In summary, while randomized evaluation is not a bulletproof strategy, the potential for biases are well known and can often be corrected. This stands in contrast to biases of most other types of studies, where the bias due to the non-random treatment assignments often cannot be signed nor estimated.

Publication bias appears to be substantial with retrospective studies; randomized evaluations can help address publication bias problems, but institutions are also needed.

Publication bias is a particularly important issue that must be addressed: available evidence suggests the publication bias problem is severe (DeLong and Lang, 1992. There is a natural tendency for positive results to receive a large amount of publicity: agencies that implement programs seek publicity for their successful projects, and academics are much more interested and able to publish positive results than modest or insignificant results. However, clearly many programs fail, and publication bias will be substantial if positive results are much more likely to be published.

Publication bias is likely to be a particular problem with retrospective studies in which, ex post, the researchers or evaluators define their own comparison group, and thus may be able to pick a variety of plausible comparison groups; in particular, researchers obtaining negative results with retrospective techniques are likely to try

different approaches, or not to publish. In the case of "natural experiments" and instrumental variable estimates, publication bias may actually more than compensate for the reduction in bias caused by the use of an instrument because these estimates tend to have larger standard errors, and researchers looking for significant results will only select large estimates. For example, Ashenfelter, Harmon and Oosterberbeek (1999) show that the there is strong evidence of publication bias of instrumental variables estimates of the returns to education: on average, the estimates with larger standard errors also tend to be larger. This accounts for most of the oft-cited result that instrumental estimates of the returns to education are higher than ordinary least squares estimates.

Hence, evaluations in which the comparison group is chosen in advance (such as randomized evaluations) can help alleviate problems of publication bias: once the work is done to conduct a prospective evaluation the results are usually documented and published even if the results suggest quite modest effects or even no effects at all. It is important to put institutions in place to ensure negative results are disseminated. Such a system is already in place for medical trial results, and creating a similar system for documenting evaluations of social programs would help to alleviate the problem of publication bias. Beyond allowing for a clearer picture of which interventions have worked and which have not, this type of institution would provide the level of transparency necessary for systematic literature reviews to be less biased in their conclusions about the efficacy of particular policies and programs.

Although any given randomized evaluation is conducted within a specific framework with unique circumstances, randomized evaluations can shed light on general issues. Without a theory of why the program has the effect it has, generalizing from one well executed randomized evaluation may be unwarranted; however, similar issues of generalizability arise no matter what evaluation technique is being used. One way to learn about generalizability is to encourage adapted replications of randomized evaluations in key domains of interests in several different settings. While it will always be possible that one program that was unsuccessful in one context would have been successful in another adapted replications, guided by a theory of why the program was effective, will go a long way towards alleviating this concerns. This is one area where international organizations, which are already present in most countries, can play a key role. Such an opportunity was seized in implementing adapted replications of PROGRESA in other Latin American countries. Encouraged by the success of

PROGRESA in Mexico, the World Bank encouraged (and financed) Mexico's neighbors to adopt similar programs. Some of these programs have included randomized evaluations (for example, the PRAF program in Honduras), and are currently being evaluated.

It is worth noting that the exogenous variation created by randomization can be used to help identify a structural model. Attanasio et al. (2001) and Berhman et al. (2002) are two examples of using this exercise in combination with the PROGRESA data to make some prediction of the possible effect of varying the schedule of transfers. For example, Attanasio et al. (2001) found that the randomized component of the PROGRESA data induced extremely useful exogenous variation that helped in the identification of a richer and more flexible structural model. These studies rest on assumptions that one is free to believe or not, but at least these studies are freed of some assumptions by the presence of this exogenous variation. The more general point is that randomized evaluations do not preclude the use of theory or assumptions: in fact, they generate data and variation that can be useful in identifying some aspects of these theories.