DOING IMPACT EVALUATION No. 6

Data for Impact Evaluation

THE WORLD BANK

Poverty Reduction and Economic Management PREM

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

Data for Impact Evaluation

October 2007

"Too much emphasis has been placed on formulating alternative econometric methods for correcting for selection bias and too little given to the quality of the underlying data. Although it is expensive, obtaining better data is the only way to solve the evaluation problem in a convincing way."
– Heckman, LaLonde, and Smith (1999) in *Handbook of Labor Economics*, Ashenfelter & Card, eds.

"If you can't be with the one you love, love the one you're with." – Stephen Stills

Acknowledgements

This document was written by Paul Wassenich, with contributions from Juan Muñoz. The authors gratefully acknowledge many suggestions which have improved it considerably, in particular from Markus Goldstein and Arianna Legovini. Any errors remain solely those of the authors.

This paper is part of a series on Doing Impact Evaluation which has been financed through grants from the Trust Fund for Environmentally and Socially Sustainable Development supported by Finland and Norway and the Bank-Netherlands Partnership Program funded by the Netherlands. The task manager of this note was Markus Goldstein.

TABLE OF CONTENTS

Introduction: Data for Impact Evaluation

Good quality data are one of the most important aspects of conducting an impact evaluation, but obtaining or generating high quality data entails trade-offs and difficulties. Compromises inevitably need to be made. Collecting new data is not always necessary for carrying out a rigorous impact evaluation, but it is often desirable where affordable. And sometimes, existing data are simply not sufficient to conduct an impact evaluation. This paper focuses on the effective and appropriate use of existing data and on issues involved in collecting new data, as well as on exploiting combinations of existing and new data.

Guide to this Document

The paper is structured as follows. Section I introduces considerations of the kind of data needed for a particular evaluation, which can help determine the best mix of existing and new data for the evaluation at hand. Section II reviews common sources of existing data used in program evaluations. Section III offers some ideas for where to look for existing data. Section IV talks about how to determine the feasibility of getting various kinds of data, existing or new, under various constraints. Section V briefly discusses steps to ensure high quality primary data or to check the quality of secondary data. Section VI discusses sampling issues, primarily for collecting new data but with applications for existing data as well.

The order of the sections roughly corresponds to the order of the steps in thinking through the data requirements of the evaluation. Before looking into what is available, the evaluator should think about what information is needed for the evaluation. This will help focus the search and efficiently assess the utility of existing data as they are identified and accessed. Second, it helps to have a general idea of what may potentially exist, to know what to look for. With that in mind, it is useful to have some ideas for some initial places to start searching for existing data. After identifying sources of existing data, it is time for a reality check as to what can really be obtained, either through accessing existing secondary data and/or by collecting new primary data, within given time and resource constraints. Before analysis is conducted, it is important to verify the quality of the data being used, to know how much stock one can put in the results. Lastly, if primary data are being collecting, getting the sampling strategy right is crucial to allow for the most rigorous methods to be used in the evaluation. Overall, since methods depend on data, parts or all of the above process may be iterated more than once as the evaluation team determines the optimal approach within the constraints they face.

Several text boxes appear in various parts of the paper to highlight specific impact evaluations that have obtained or used data in creative ways. These serve to illustrate some of the concepts presented in the main text, as well as to provide examples that may help evaluators to think about how they can approach gathering and generating data for their studies. In particular, they spotlight studies which have made good use of secondary data to evaluate development programs.

I. Determining What Data Are Needed

Data and methods are inextricably linked. So the kind of data an evaluator needs depends on the questions she wants to answer and the methods she hopes to use to answer them. Other papers in the Doing Impact Evaluation Series[1] cover topics related to impact evaluation methods in more depth. This paper focuses on the data side and mentions methods in passing as they relate to data issues. Four principal considerations in conceptualizing the ideal dataset for an evaluation are:

A. what indicators are needed,
B. at which level of observation,
C. collected at some appropriate point(s) in time, and
D. whether existing data sources contain some or all of these necessary indicators for the appropriate level of observation at the right time(s).

It is often tempting—particularly under time constraints—to think "in general" about the kind of information the evaluation team will want to have and proceed quickly to obtaining the data or planning a survey. However, it is worth investing the effort up front to think carefully about the model behind the program being evaluated and exactly what information will be needed to test the hypotheses that the model implies. Otherwise, the data gathering process may leave out information needed for the evaluation.

A. Indicators

First, an evaluator needs to determine what information she will need in order to test her hypotheses regarding program impacts. This includes determining the specific indicators needed on:

1. impacts of interest;
2. outcomes of interest;
3. outputs, typically program participation or "treatment dosage" received; and
4. external factors which could also influence these outcomes.

In the language of regression analysis, these are called:

1. dependent, or left-hand-side (or LHS), variables;
2. dependent, or left-hand-side (or LHS), variables (since impacts and outcomes are treated the same way in regression analysis);
3. independent, or right-hand-side (RHS), variables indicating treatment (or sometimes even input indicators, since project outputs correspond to causal inputs in the econometric model); and
4. control variables, also referred to as independent or right-hand-side variables.

An evaluation requires indicators in *all* of these categories to generate reliable results. Analysis based on outcome and impact indicators without appropriate control

[1] http://go.worldbank.org/169GZ6W820.

variables is almost certain to yield biased results. Worse yet, it is impossible in such cases to determine the direction of bias. One only knows that the answers are "wrong," without knowing in what way they are wrong. This concern (usually referred to as "omitted variable bias" or "model specification error") has always occupied a central place in econometrics.[2]

Obviously, the specific indicators needed will depend entirely on the nature of the project being evaluated: its outputs, its objectives, its intended recipients. A discussion of this topic in general terms is therefore necessarily limited. Nevertheless, to make the discussion a little more concrete, some common examples of each category of indicator include:

- income, wealth, mother and child mortality, educational attainment;
- employment status, pre- and post-natal medical consultations, school attendance;
- amount of money invested per community, or a binary indicator of participation (one for participants and zero for non-participants); and
- mother's educational attainment, household wealth, distance to existing infrastructure, etc.

Note that some indicators (e.g., wealth, educational attainment) can serve different roles depending on the nature of the analysis: sometimes serving as an impact or outcome indicator; other times as a control variable.

To help the evaluator determine the complete set of indicators that are needed for the evaluation, the project and evaluation teams should map out the development model underlying the project, identifying the links from outputs to development outcomes and impacts. In addition to project outputs and final outcomes, this should include mapping out assumptions and intermediate outcomes, so that indicators to track changes along each link of all the chains of causation can be identified.

Ultimately, an impact evaluation should strive to identify not just *if* but also *how* a project delivers its results. Identifying where a chain of causation breaks down in the case of unexpected results is important in determining if the model does not accurately reflect reality, or if implementation deviated from the model. Making these determinations can help improve the model upon which the intervention is based, by providing specific points to consider in designing, modifying, and implementing the intervention in the future. Identifying where assumptions or intermediate outcomes may be violated and yet still arrive at intended outcomes can also find weaknesses in the model that merit further research. As is discussed in more detail below, this suggests that administrative data collected for monitoring purposes has a key role to play in impact evaluation, regardless of what other data may be employed.

It is worth mentioning here that when survey questionnaires are being designed for primary data collection, they should strive to generate indicators in the same way as existing questionnaires in the country have done in the past, to improve comparability.

[2] See, in particular, Leamer (1983), and LaLonde (1986), or any econometrics textbook.

That is to say, if a new questionnaire is intended to generate indicators of education, it makes sense to copy the education questions from the census or national household survey, unless these are found to be lacking in some way. (See also section V.A.2 below on how questionnaire design influences data quality, including how not just the wording, but also the sequencing, of questions can generate different responses.)

B. Level of Observation

Another consideration in determining what data are needed for an impact evaluation is the level of observation. The appropriate level of observation for a dataset depends on the type of intervention and its expected benefits. Many evaluations may require observations at more than one level. For instance, a project implemented at the community level to improve households' income may require community-level observations to model program selection and placement but household data to measure income and/or consumption. Similarly, facility data might be needed for modeling program placement (in the case of an intervention related to, say, schools or clinics), and potentially even for some outcomes (like teacher or doctor attendance), while individual-level data on test scores or disease incidence is used as impact indicators.

Efficiency considerations also come into play when thinking about the level of observation. For instance, if data on infrastructure can be reliably collected at the community level, there is clearly no point in asking households or individuals about the presence of that infrastructure in their community. However, measuring access to existing infrastructure may require data at the household or individual level. Likewise, community averages may be of interest (e.g., village poverty rate), which can only be calculated from data at a more disaggregated level (household in this case). For examples of studies which have combined data on different levels of observation, see Boxes 1, 3, and 5 below.

This section briefly discusses five of the most common levels of observation that are commonly used in program evaluations.

1. ***Individual:*** For projects which are implemented at an individual level, data collected at this level may be sufficient for an evaluation. An example of this is labor-training programs which are intended to decrease participants' time out of work and to raise their income. Indicators of outcomes (time searching for a job, daily wage), outputs (participation, days of training received), and controls (previous education, work experience, etc.) can all be observed at the individual level.

2. ***Household:*** Many projects are intended to engender benefits at the household level. This requires data at the household level to measure those outcomes. Typically, this involves observing indicators that apply to the entire household, such as housing characteristics and asset ownership, as well as indicators for each individual, such as age, educational attainment, employment status and earnings, etc.

The researcher often needs to think carefully about the economically appropriate level of observation between the household and the individual. Households may benefit from some economies of scale, or have certain complementarities or substitutions in consumption that are not evident at the individual level. Conversely, intra-household allocation of resources or benefits may be of interest in the evaluation, in which case, household-level data would hide effects that could only be observed using data on individuals within households.

Box 1: Combining Existing Household and Municipality-Level Data to Evaluate Water Provision Privatization

Galiani, Gertler, and Schargrodsky (2005) studied the impacts of the privatization of water provision on child mortality in Argentina during the 1990s based entirely on secondary data compiled from several sources.

For the primary impact indicator, data on child mortality rates by municipality and by cause of death were compiled from vital statistics maintained by the Ministry of Health. The main outcome indicator was the proportion of households connected to the water network, which was obtained from the 1991 Population and Housing Census and the 1997 Social Development Survey, a large-scale household survey, both of which included the same question about household access to water. The main indicator of program inputs was from data on municipal water provision from the national sanitation authority. Control variables on municipal characteristics were compiled from the 1991 census data. Since the intervention took place at the municipal level, the analysis was carried out using data at the level of the municipality, either directly obtained or calculated from census data on households in the municipality.

3. ***Community/Village***: Many projects are implemented at the community level and thus may require data at this level for an evaluation. For instance, if a program delivers small-scale infrastructure to poor communities, observations of (pre-) existing infrastructure are needed at the community level to model selection of beneficiary areas.

Box 2: Compiling Existing Municipality Data to Evaluate an Anti-Corruption Program

Ferraz and Finan (2007) studied the impacts on 2004 electoral outcomes from an anti-corruption program of audits of municipal expenditures initiated in Brazil in 2003, using secondary data compiled from several different sources, including data they coded up themselves using publicly available audit reports.

For impact indicators, they used data on election results available from the electoral tribunal. For indicators on outcomes of the program, they coded up data themselves from audit reports available on the program's public website. For the program input variable, they used the timing of the public release of the audit findings for each municipality relative to the 2004 mayoral elections. For control variables, they obtained data from the electoral tribunal on candidate characteristics, as well as data from the national statistical institute (from the 2000 population census and a 1999 survey of municipalities) on municipality characteristics.

4. ***Facility/Infrastructure:*** Projects may also be implemented at the level of a facility or other infrastructure, such as a school, health center, road, etc. For an evaluation of such projects, the evaluator may need to observe some characteristics at this level, for control or outcome variables.

5. ***Firm/Enterprise:*** For private sector development, financial sector development, and some infrastructure projects, data on private companies are important to the evaluation, for indicators on outcomes and impacts, as well as control variables. See Box 4, for instance.

C. Timing of Data Collection

At a minimum, indicators need to exist for one period in time after a project has been implemented, with both treatment and comparison observations (that is, with some observations that participated in the program and others that did not). For more rigorous evaluation methods, it is necessary to have data from at least two points in time, before and after project implementation, on the same group of observations. This allows the comparison of progress over time between program participants and non-participants. Repeated observations over time can allow the estimation of trends before, during, and after project execution. Of course, one cannot recreate data in the past (and retrospective questioning is often considered unreliable), so depending on when the evaluation is planned relative to the project, the evaluator may have to content himself with what is available from the past and what can be collected in the future.

The timing of data collection for the purposes of impact evaluation is more sensitive relative to other common motivations for collecting data. It is obvious that baseline data collected after program activities have begun are not really baseline data. But it can also be problematic if baseline data are collected too far in advance of program implementation, such that they do not provide an accurate representation of the situation at the time the program began. Likewise, ex post data should not be collected too far from the time that expected benefits are likely to be perceived, or the passage of time may muddy analytical results. In some cases, program impacts may mostly take place in the immediate and short term, in which case ex post data should be collected soon after project closure. In other cases, impacts may not be fully realized until the medium or long term, in which case it may be necessary to collect a round of ex post data several years after project closure in order to calculate a true measure of these program impacts.

The value of early planning for an impact evaluation cannot be overemphasized. It allows more strategic and more complete data generation. Related to this, it allows more rigorous evaluation methods to be employed.[3] Collecting relevant baseline data depends critically on planning the evaluation during the program design phase. This is true regardless of whether baseline data are collected specifically for the purposes of the evaluation or whether the evaluation intends to piggyback on some other data collection that happens to come at the appropriate time.[4]

[3] See section IV.A below for more on the virtues of planning an impact evaluation as early as possible in the project design process.

[4] See section IV.B below on using existing data and piggybacking on other data collection efforts.

D. Primary vs. Secondary Data [5]

Once the evaluator has determined what kind of data she needs, including which indicators and what level of observation and which points in time are required, she can search for existing data that may satisfy these requirements.[6] Sometimes an evaluator can find existing data that meet the needs to address some or all of the research questions in the evaluation and only needs to obtain that data and conduct the analysis. (This is more common than one might initially suppose.[7]) Other times, existing data have some of the information needed but must be supplemented with new data before the analysis can be done. (This is quite common.) And at times, no data exist with information that can be brought to bear on the evaluation, and the evaluation must rely entirely on data newly collected for the purposes of the evaluation. (This is indeed rare.) As is discussed in the subsection on administrative data below, all projects should have monitoring data that will be relevant for the purposes of impact evaluation.

The decision regarding the balance of existing and new data essentially rests on two factors: how far available data can go toward answering the questions one wants to evaluate and what resources are available for collecting new data. A third consideration is whether existing plans for future data collection by other parties can be exploited to generate new data that are needed. Economies of scale certainly exist in data collection, and evaluators can sometimes piggyback on other data collection efforts to reduce their costs. To generate necessary indicators, it is usually possible to add new questions or modules to a planned survey questionnaire, if one or both of the collaborating parties can cover the additional costs this implies. This is not without certain potential trade-offs: *sampling*, *timing*, and *quality* being three points to consider.

If the *sampling frame* of a survey that is being planned does not sufficiently overlap with the sample needed for the evaluation, it may not be worth joining forces with that particular survey. This could happen if the data needed for the evaluation should focus on certain geographic areas that are not well represented in the sampling frame. It could also happen if the evaluation intends to focus on a minority group—such as families with infants, the elderly, a particular ethnic group, or disabled persons—where a survey representative of the general population will generate very few observations of interest.

Second, if the *timing* of the other survey is too early or too late for the purposes of the evaluation, it may not generate data useful for the evaluation. Section I.C above discussed the importance of the timing of data collection relative to the implementation of the intervention.

Third, if an evaluation team has doubts about the *quality* of secondary data, it is well worth investing the effort in doing some quality checks. An evaluation carried out

[5] By primary data, we refer to data newly collected for the purpose of the evaluation; and by secondary data, we mean existing data, or data collected by agents outside the evaluation team.
[6] Sections II and III below talk about finding existing data to use in an evaluation.
[7] The studies highlighted in Boxes 1 and 2 are examples of evaluations carried out entirely with secondary data.

using poor quality data will lack credibility. More discussion on checking data quality can be found in section V below.

Sections III.B and IV.B below contain further discussion of combining data from different sources for an impact evaluation. This includes cases of using existing data in conjunction with new primary data and piggybacking on other data collection efforts.

II. Potential Sources of Data

Once the evaluator has determined the characteristics of the information he needs, it is time to see how much of it already exists. This section briefly describes various types of data more frequently used in impact evaluations, and the following section discusses some tips for where to start looking for data when one is not sure what exists.

A. Administrative Data

Administrative data are collected and maintained by various government entities and by authorities who implement and manage projects and programs. These data are frequently useful for monitoring where a program took place or who participated. They may also contain information on the amount spent in various locales. They are sometimes useful for obtaining other indicators, such as school attendance, medical consultations, infrastructure maintenance, etc. But administrative records in some countries are notoriously inaccurate, so it is wise to verify their reliability before using the data in an evaluation.[8]

In particular, administrative data for the program being evaluated can be very important, especially project monitoring data. They are often necessary for the identification of treatment and comparison observations. They can also play a crucial role in providing a precise characterization of the nature of the intervention being evaluated, including how its design and implementation may have varied by location and over time. Program design and structure can certainly change during execution, in response to any number of factors. Administrative data are needed to understand exactly how and when a program was altered. Furthermore, program implementation may deviate from program design, either deliberately or in an ad hoc manner. This may not always be recorded in any official documentation related to the program. It then falls to administrative data to provide the needed information on how the intervention actually took place in practice. This includes whether it was implemented consistently across units, or whether implementers had discretion in how they carried out their activities (and if so, how much latitude was exercised).

Furthermore, some changes in the practice of collecting project monitoring data can vastly improve their usefulness for impact evaluation. A primary example would be to collect data not just for participants but also for those who expressed interest but could not participate. For instance, communities that applied for social funds, but whose project

[8] See section V.C below on checking data quality.

was not approved, or groups or individuals who were not approved for a microfinance loan, or individuals deemed ineligible for a labor training program, etc. Such data could be useful in two ways. One, the latter group may form a natural comparison group. Two, these data may allow for better understanding and characterization of project selection and placement in practice.

B. Household Survey Data

National household surveys are periodically conducted in many countries. These include multi-topic household surveys, which can cover a wide array of information on housing characteristics, household consumption and wealth, and individual employment, education, and health indicators. Labor force surveys are more restricted in topical scope and sometimes cover only urban areas. Two commonly used multi-topic household surveys are the Living Standards Measurement Surveys (LSMS)[9] and Demographic and Health Surveys (DHS),[10] which have been conducted in many developing countries, sometimes including several rounds of data collection over the past 25 years or so. LSMS are typically collected by the national statistical agency and may go by different names in different countries.

Household surveys may provide rich data on household and individual characteristics. However, two complications frequently arise. The first common challenge in using household survey data for impact evaluations is reliable identification of beneficiaries. If there is no indicator in the questionnaire regarding the program in question, the fallback option is often to use geographic indicators, considering those in areas where the project took place to be beneficiaries and those outside areas affected by the project to be non-beneficiaries. Obviously, this only works for non-exclusionary interventions: village-level projects, for instance. If this can be done, the second obstacle which the researcher may face is limited overlap between the areas sampled for the survey and the areas covered by the project. The sample of beneficiaries in such cases may be quite small, which can reduce the precision of the quantitative analysis.[11]

C. Census Data

Most countries conduct a population and housing census every ten years or so, and many also conduct an agricultural census, usually with less frequency. The advantage of census data is that they aim to cover the entire population, so there are data for virtually every potential treatment and comparison observation. However, as with household survey data, identifying beneficiaries can be a challenge. The drawback of census data is that they typically contain a limited number of indicators, often limiting their value for an impact evaluation.

Combining census data and national household survey data, researchers can generate very detailed poverty maps and small area estimation of poverty. This allows for more heterogeneity within areas to be measured. This is done by exploiting detailed data

[9] www.worldbank.org/lsms
[10] www.measuredhs.com
[11] See section VI on sampling below.

on consumption, income and/or wealth from the household survey, and the comprehensive coverage of the census, via common indicators in both.[12]

Because of their comprehensive coverage, census data can be useful in the randomization of treatment and control areas for an evaluation using experimental design methods.[13] In particular, if an intervention is to be targeted to areas with certain characteristics, and a targeted or stratified sample[14] is needed for the evaluation, the census can be used to identify areas with the desired characteristics, from which treatment and control areas can be randomly assigned.

Box 3: Combining Census, Household Survey, and Administrative Data to Evaluate a School Construction Program

Duflo (2001) combined several sources of secondary data for her study on the impacts of school construction in Indonesia. The 1995 national household survey was used for individual-level data on wages and education for men born 1950 and 1972. The younger were of school age during the reform period, 1973-1979, when 61,000 primary schools were constructed; the elder were too old to benefit from new primary schools. These data were linked with district-level data on the number of new schools constructed during the 1973-1979 period, by matching individuals with their district of birth. "Treatment dosage" was determined by the number of new schools built in the district and the individual's age at the time the program began. This is further supplemented with individual-level data on incomes from the 1993 SUSENAS survey, district-level averages from the 1971 census of the proportion of population in school, and "baseline" data from the Ministry of Education on primary school enrollment in 1973.

D. Facility Survey Data

Facility surveys are conducted to collect data at the level of service provision, e.g. a school or a health clinic. In many countries, the ministries of health and education compile information on clinics and schools on a regular basis. Additionally, these data may in some cases be collected through stand-alone surveys or in conjunction with other surveys, such as a household survey or a specialized survey. Data from facility surveys are often most useful in analysis for providing control variables on infrastructure characteristics, e.g. student-teacher ratio for each school. However, in some cases, facility-level indicators may be outcomes of interest (e.g., again, student-teacher ratios), or some other indicators that try to capture quality of service provision (e.g., teacher attendance rates; presence of a doctor, nurse, and/or pharmacist at a health center, etc.).

E. Industry Data

Some government ministries or other entities maintain data on industries related to their mandate. For instance, data on agricultural commodity prices and quantities traded are often maintained by the Ministry of Agriculture, often for sub-national markets that allow researchers to observe price differentials within a country. General price and

[12] For more details on small area poverty estimation, see http://go.worldbank.org/0IP8E8O1B0.
[13] See Ravallion (2006) for impact evaluation methods.
[14] See Section VI on sampling.

other market data may be collected on an ongoing basis by any of several agencies: the central bank, the finance ministry, or the statistical agency. Non-governmental industry bodies may also collect similar data.

F. Specialized Survey Data

A specialized survey is one that is collected for a specific purpose, often research on a particular topic. They are frequently household surveys but may be facility, firm, or individual level surveys or some combination thereof. Many take modules from the existing national household survey questionnaire and add questions on topics of interest which the national household survey does not include. Coverage of specialized surveys can be quite limited, sometimes resulting in limited or no overlap with program areas. Nevertheless, if the evaluation team can find existing data from a specialized survey on a topic related to the evaluation, these datasets can provide a rich collection of relevant indicators.

G. GIS / GPS Data

Geographic Information Systems (GIS) data, and Global Positioning System (GPS) data in particular, can provide precise indicators on the locations of various infrastructure and geographic terrain features. For instance, GIS data may have indicators on the locations of schools, hospitals, public telephones, roads, rivers, etc., as well as data on elevation, contour, terrain type, etc. In evaluations where spatial relationships are important, GIS data can supplement other data to enrich the analysis. For instance, GPS readings could be recorded for each household in a survey, and these data could be used with existing or new GPS data to estimate travel times to the nearest market, police station, school, health post, etc. Recording GPS coordinates for each dwelling in household surveys also greatly facilitates the construction of panel datasets by making it much easier to locate the dwellings to include them in a later round sample.[15]

III. Tips on Locating Useful Data

Before getting into recommendations regarding where to look for existing data, it is worth mentioning an important recommendation regarding general data management here.

It is always a good idea for project teams and evaluation teams to keep back-up electronic copies of all raw data. This includes even census and national household survey data. Original data have been lost by the agencies that collected it—even national statistical agencies—with more frequency than one might suppose. In some cases, it is not clear who has the responsibility to maintain and archive data, and thus nobody does.

Even when data are properly preserved, the agency that does so may not be inclined to share it at a later time. For example, data collected as part of a project that received Bank funding may be archived by a line ministry for years. While the ministry

[15] See section VI.D on panel data below.

may have been inclined to share the data with Bank researchers during the project's lifetime, the political and bureaucratic staff change over time, and the new regime may no longer be willing to share the data after a few years have passed. Imagine the following (not entirely fictional) scenario. At the project design phase, a project team was interested in conducting a rigorous evaluation, and through numerous discussions, they convinced the government authorities of the value of learning serious lessons from the experience of the project, so as to improve their service provision in the future. A line item specifically for impact evaluation, including data collection, was written into the loan. The project execution lasted eight years, and another researcher was interested in conducting an ex post impact evaluation two years after the project had closed. The government staff involved in the project had changed. The new staff had not been not party to the initial discussions regarding the value of an impact evaluation and so were not as inclined to share the data as those involved at the beginning of the project had been.

Moreover, different agents within the government may well have different opinions of the value of an impact evaluation. From a political economy standpoint, the finance ministry may be very interested in conducting a rigorous impact evaluation, since its results may inform their decisions regarding resource allocations. However, relevant line ministries may see an impact evaluation as a risk, since they could stand to lose funding if the results are disappointing. The germane point for the evaluation team is that seeking out and building up coalitions within the government to support an impact evaluation can pay off in terms of making the evaluation happen and in terms of accessing data.

To judge how useful a particular dataset is likely to be, the evaluator needs to know what indicators it contains, as discussed in section I above. The simplest way to do this is to request a copy of the questionnaire used to collect the data. Questionnaires are usually shared readily, while getting access to full datasets may require more work. Going through a questionnaire first may save the evaluation team the hassle of requesting the full dataset only to find it does not have any indicators they can use.

A. Places to Look

To use existing data, a researcher must first find out what data exist and then get access to those sources. A good place to start is the *International Household Survey Network* (IHSN) at http://www.internationalsurveynetwork.org/home. This has two important resources: (1) the Central Survey and Census Catalog and (2) the Microdata Management Toolkit.

The Central Survey and Census Catalog is a listing of existing, as well as planned and on-going, data sources organized by country. This is a good place to start to see what may be available, but it is not intended to be an exhaustive list for any country. It does not, for instance, include administrative data or specialized surveys.

The Microdata Management Toolkit is intended for data producers (e.g., statistical institutes, researchers collecting primary data, etc.) to help them organize and disseminate data in a consistent and high quality manner. This serves two purposes. One

is to help data producers verify their datasets are properly documented for dissemination, by providing quality assurance checks in the toolkit. Another is to facilitate access to existing data by providing data producers with software to easily convert data into an accessible, "standardized" format, which can then be used in various statistical software packages. For impact evaluations that include primary data collection, a team must be contracted to go out and collect the data in the field. The Microdata Management Toolkit can help the field researchers organize, compile, and clean the data and can help the analysts check the quality of the dataset.

The World Bank has developed the *Development Data Platform* (DDP), which provides information on existing household survey datasets and their characteristics. For World Bank internal use only, the Development Data Platform (DDP) provides access and basic analytical tools for both time series (macro) and survey (micro) data on a wide range of development topics; and includes metadata, documentation and related datasets. DDP includes features with which users can prepare and publish web reports, charts, and maps.

A very useful webpage can be found by going to www.worldbank.org/impactevaluation, then clicking on "Data and Tools" in the menu on the left, then clicking "Accessing Surveys."[16] This page has links to several lists of data sources, including the IHSN, the DPP, the Africa Household Survey Databank, the MECOVI Household Surveys Initiative for Latin America and the Caribbean, and World Bank Research Department datasets.

Contacting the statistical agency in the country where the evaluation is taking place is a good idea. They can inform the evaluator of existing and sometimes upcoming surveys in which they have been involved. But they may not know of surveys that were run outside their purview, so this is not a comprehensive source of information.

A survey of the empirical literature on topics related to the evaluation sometimes reveals studies carried out in the same country as the project being evaluated. Contacting the authors of those studies can sometimes lead to data sources one might not find otherwise, often specialized surveys related to the topic of the evaluation that include indicators not often found in other data sources.

Talking to the project implementation agency is essential for any evaluation, but it is worth asking them to identify all sources of information related to the project and where they would turn to get the information you are looking for. When talking with people without a background in quantitative research, the words "data" and "statistics" can take on any number of meanings, so it pays to spell out what you need in detail. Many people hear "data" and think "statistical tables" or "output reports," so it behooves researchers to use less ambiguous terms like "microdata," "raw data," or "unit record data."

[16] The direct URL is http://go.worldbank.org/B50PMCIUV0.

B. Combining Data from Different Sources

Using data from more than one source is common in impact evaluations. Different approaches can be taken to integrate data from two or more sources. One is to merge the separate datasets into one dataset which is used for analysis. Another is to conduct the analysis in stages, each stage using a different dataset, where the results from an earlier stage of analysis are used as inputs for a later stage. For evaluations that investigate multiple topics, another alternative could be to apply the best analytical method possible to each question, given the availability of the necessary indicators in the various datasets. Each of these approaches is discussed in this section.

Merging data from different sources into one dataset relies on exploiting shared coding across datasets to match observations. For instance, if two or more datasets contain different indicators for overlapping sets of households, merging them puts all the indicators available for each household together in one dataset that can then be used for analysis. As GPS locations are included in more datasets, these will become a useful indicator for matching observations across datasets.

One case of merging datasets is that of combining rounds of panel data, where households or individuals (or ideally both) are given identification codes to facilitate matching observations across time periods. However, if the data collection were not carefully implemented, different rounds of panel data may be difficult to match. Even though common codes are used to identify households over time, matching individuals within households from one time period to the next can be tricky. Codes for individuals from the first round of data collection are not typically maintained in the latter round, in part because household composition inevitably changes over time. Matching individuals within a household using age and gender is one option, but implementing this in software like Stata requires complicated algorithms.[17]

Furthermore, matching household ID codes across datasets can sometimes be very frustrating in practice, not just because of its intrinsic difficulty, but as a result of poor quality fieldwork and data management. For instance, the Nepal Census of 1991 never could match household and individual data, which were separated to speed-up data entry without sufficient attention to maintaining ID codes on all the completed paper forms.

Box 4: Merging Multiple Dataset Using Firm ID Codes to Evaluate Promotion of Firm-Level Research and Development

De Negri, Lemos, and de Negri (2006a, 2006b) use a database compiled by IPEA (Institute for Applied Economics Research) to study the effects of two firm innovation programs in Brazil. IPEA matched observations by unique firm identification codes from six separate datasets (one each from the Ministry of Labor, the Ministry of Industrial Development and Foreign Trade, the Central Bank, and the National Institute of Industrial Property, and two from the statistical institute). The combined dataset contains indicators on the outcome of firms' research and development expenditures and on the impacts of firms' patents, productivity, and profitability.

[17] See section VI.D below for more discussion of panel data.

In other circumstances, merging datasets often becomes a question of what is pragmatically feasible rather than what is ideal for the evaluation design. An evaluator would like to match observations at the level used in the evaluation, typically households or individuals. However, because individuals or households are rarely designated by standardized codes across datasets, combining data is typically done at a more aggregated level where standardized codes do exist.[18] This is most often done using geographically-defined codes, perhaps census tracts or some larger administrative unit. Yet even geographic codes are not always commonly defined across agencies which collect data. For instance, the statistical institute may have one set of codes for census tracts, while the agricultural ministry has another set for land units whose boundaries may not correspond to those of the census tracts. In such cases, the evaluation team must find a way to generate common coding from the two datasets—e.g., which agricultural parcels fall within certain census tracts—in order to merge them.

GPS coding can overcome such hurdles, and can even serve as a cross-check where common administrative codes may exist. Further, GPS data could be used to match households, and/or land parcels, across datasets which have included GPS locations. As such, including GPS data in primary data collection is a good practice where it is feasible, both in terms of being able to match primary data to existing secondary data sources (both past and future) and in terms of making primary data more accessible and useful for other researchers who might use them as secondary data in the future. However, including GPS locations in publicly available data could jeopardize the confidentiality under which they were collected. It is therefore necessary to treat these fields as personal identification that is not included in the published version of a dataset.

Where shared coding is not available, alternative methods of matching may be possible, but they are frequently cumbersome to implement. Using address, telephone number, or name as a proxy for identification codes to match households or individuals is sometimes possible, but formatting and spelling need to be exactly the same for software to recognize the match. In some cases this problem can be circumvented using intelligent computer pattern-matching algorithms (so-called Gestalt pattern matching algorithms).

Another approach to combining data from different sources for an evaluation is to conduct the analysis in stages, each stage using the appropriate dataset. Say, for instance, that one wants to evaluate a project intended to provide needed infrastructure to villages lacking primary schools, public health clinics, and/or potable water sources. Available data consist of a community-level infrastructure inventory collected prior to project implementation, which were used to identify and select villages lacking critical infrastructure, and a household survey collected after the project closed. In this scenario, an evaluator could potentially use the community-level data to model project selection and placement for propensity score matching of treatment and comparison villages. The household survey data could then be used to compare outcomes for similar households in matched communities.

[18] Collecting national identification numbers would work very well for matching individuals across data sources but could jeopardize the confidentiality under which much data is collected. Hence it is rarely done and virtually never available.

Box 5: Using Community Data for Matching and Household Data for Impact Variables to Evaluate an Infrastructure Rehabilitation Program

Lokshin and Yemtsov (2005) employ this approach in an evaluation of the impacts of an infrastructure rehabilitation program in rural Georgia during 1998-2000. They use retrospective community-level data from the 2002 Rural Community Infrastructure Survey for propensity score matching of pairs of similar treated and comparison villages. They then use longitudinal household survey data available for 1996 onwards from the ongoing multi-topic Survey of Georgian Households for impact indicators in difference-in-difference estimation of impacts.

A third approach to using separate datasets in an evaluation is to use different methods to address different questions, as data availability allows. The evaluator should use most rigorous methods possible for each question in the evaluation, as determined by data availability. Other questions, for which indicators are not present in all the data sources, can be analyzed using less rigorous econometric methods. Take, for example, an evaluation of a social fund, with population census data collected before the project began and an extensive household survey collected after the project closed. For the few outcome indicators present in both the census and the survey, difference-in-difference estimation could be possible. For the majority of outcome indicators present in the survey but not in the census, only cross-sectional analysis is possible.

A good practice in such a scenario is to use both methods on the questions which are also addressed using the more rigorous methods, in order to compare the results and see if the weaker methods yield similar results to the superior methods. This makes for a decent robustness check on the weaker methods. If the results are not consistent, the evaluator must carefully consider whether the results from the weaker methods may be biased for all the questions they are used to address.

IV. Determining What Data Can Be Feasibly Obtained

Having conceptualized the ideal methods and dataset for the impact evaluation, and having scouted out the existing secondary data, the evaluation team must determine how close to the ideal they can feasibly get, through accessing existing secondary data and collecting new primary data. This largely depends on three factors: timing, identifying and accessing useful existing data, and the resources the evaluation team has available to use for collecting new data as needed.

A. Timing

The timing of evaluation design relative to project design and implementation is the most important of these factors influencing evaluation quality. This timing falls into three discrete categories: before, during, or after project implementation. The earlier one starts, the more flexibility one has in terms of evaluation design, including data collection. Some of the most rigorous evaluation methods, such as random assignment to treatment, cannot be used if project implementation has already begun. Likewise, primary

baseline data can only be collected before project activities actually begin. (Once the well has been dug, the "baseline" period has already passed.) Otherwise, baseline data must be reconstructed from existing data, if secondary data can be found to fill this need. (See Box 3 and Box 5 above for examples of this.)

Nevertheless, even when a project is underway or already completed, it is still possible to design and carry out a decent impact evaluation. Being in this (all too common) situation does, however, make the other two factors (existing data and resources) all the more essential. Existing data may allow the evaluator to reconstruct baseline data or to match treatment and comparison observations based on characteristics prior to the intervention or even based on time-invariant characteristics, if the secondary data were collected during or after project implementation. Having more resources available will open the possibility of collecting primary ex post data to conduct at least cross-sectional analysis. Combining existing data and new data may allow some more rigorous methods to be used, e.g. difference-in-differences estimation.

Early evaluation planning may also permit piggybacking on pre-existing plans for data collection (see the following section). Even when a project has already ended, having some flexibility in the due date for the final evaluation report may allow piggybacking for the purposes of ex post cross-sectional data.

Piggybacking can even be used to construct panel data, by inserting questions from a prior survey into an upcoming survey and by adding observations from a prior survey to the sample for an upcoming survey. When copying questions or modules from prior surveys to construct panels, the original wording and formatting must be adhered to, or the responses from the two rounds will not be comparable. This is not to say that the questionnaire has to be exactly the same. Questions and sections can be added or dropped. But the indicators the researcher wants to compare over time must be collected in a way that ensures that they are really the same indicators at different points in time. Since the wording and the sequencing of questions can strongly influence the responses given, these must be consistent for results to be comparable. (See additional discussion of piggybacking on other data collection efforts in the following subsection and further discussion regarding panel data in section VI.D below.)

B. Existing Data and Piggybacking

Finding and accessing existing data can help regardless of the timing of the evaluation, but it becomes more critical the later an evaluation begins, since it is impossible to generate new data in the past. Nevertheless, even well-funded evaluations that are planned prior to project implementation can still benefit identifying and exploiting useful secondary data. There is no sense in duplicating data collection that has already been done or is already being done.

Many issues related to identifying and accessing existing data have already been discussed in previous sections, but one key issue remains: piggybacking. This can be very effective for stretching evaluation budgets. Leveraging existing data collection efforts for the purposes of the evaluation can help the researchers get useful data for a fraction of the

cost of collecting them themselves. If the central statistical agency is already planning to do a nationwide household survey close to the time the evaluation needs data collected, why not try to combine forces? Of course, as with all data collection efforts, quality control is essential. (See Section V below.) When collaborating with other groups to collect data, it is important to verify that the data collected will be of sufficiently high quality for the intended evaluation. Using data that are not generally regarded as credible will result in an evaluation that is similarly regarded.

Piggybacking can take two forms, one or both of which may be desirable. One is adding questions or modules to existing survey questionnaires. Many countries are willing to do this if the extra cost is covered. Some statistical institutes even have a fixed pricing structure based on the amount of content added to the questionnaire.

The other is to augment the sample to include observations needed for the evaluation. A common situation is to oversample project areas to generate a sufficient number of treatment observations for a small or targeted program, relying on a sample of the general population for comparison observations. Again, many agencies are willing to administer their survey beyond the selected sample, provided the additional cost is covered. Paying the statistical agency for the marginal costs of increasing their sample by 1000 will almost always be cheaper than going out and running a 6000-household survey directly, and the evaluation team ends up with equally useful data either way. Using existing data for comparison observations and collecting one's own data on treatment observations (using the same questionnaire) is another variation on this form of piggybacking. Obviously, the timing of the data collection should be as close as possible in such cases, to avoid factors like seasonal effects or macroeconomic changes over time confounding the comparison.

Many statistical agencies are justifiably concerned about maintaining the confidentiality of the data they collect. If people perceive (rightly or wrongly) that their information is being shared in a way that jeopardizes their privacy, they may refuse to participate in future surveys, which would make it very difficult for the statistical agency to comply with its mandate. So when collaborating with outside partners to collect data, it is important to be clear up front about exactly what will be shared and delivered and the exact nature of the ownership and access rights for the data. Some agencies may only allow access to data on computers in their home offices, and only the results of econometric analysis may be taken out. In other cases, they may release data with certain variables dropped to safeguard privacy. Sometimes, these are trivial for the research, as with names; but sometimes, this can be problematic, as with professions. The evaluator must decide if the specific arrangements with her partners are going to be acceptable. Furthermore, there is a growing norm to make the data for evaluations publicly available where feasible, so it is good to express this desire up front. Sometimes, it may not be practicable or partners may simply refuse, but if some advance planning can make public sharing of the data possible, it is worth asking early.

While on the subject of working with government statistical agencies, it bears mentioning here that the evaluation team cannot contract out the study design to the statistical agency, even if they are hired to run a survey. The evaluation team should

certainly consult with them where relevant and benefit from their expertise. But the evaluator knows the motivation of the research best and hence is best placed to manage the details regarding survey questionnaire design, sampling strategy, etc. Once the evaluation team has determined exactly what indicators they want to collect, the statistical agency can be much more effective in providing expert advice on how to phrase questions, on how to structure the questionnaire, and on how to run the survey in general. Similarly, once the evaluator has determined the parameters of the needed sample, the statistical agency is probably the best suited institution to consult regarding viable sample designs, and the logistics of constructing a sample that adheres to the design the evaluator needs.

C. Resources

Resources (basically money, but also time) can help improve the prospects of having useful data for your evaluation, especially when the collection of primary data is desired or needed, and this may be more critical if existing data are scarce and/or the timing of the evaluation is late relative to the implementation of the project. Primary data collection can be fairly costly, and it is important to consider whether the evaluation budget is sufficient to carry out the planned data collection and analysis. (See, in particular, section VI.C below on sample size and statistical power.) Likewise, if the evaluation results need to be produced quickly, it may not be possible to collect and analyze primary data in time. Grosh & Muñoz (1996), in chapter 8, offer practical advice on developing a budget and work program for running a survey. Bamberger (2006) also gives guidelines for conducting evaluations with constrained resources.

V. Data Quality

Whether using primary or secondary data, or both, it is important to verify the quality of data. Conducting analysis on data full of errors and inaccurate information is, at best, a waste of time, and, insomuch as policy makers take evaluation results seriously, irresponsible. This section discusses issues which affect data quality. While the presentation here focuses mostly on primary data, the set of issues is largely the same when using secondary data. The main difference is that one is able to take active steps to address them in the case of primary data collection, while with secondary data, it is more a question of understanding how these issues were dealt with and potentially how to compensate for any problems. Subsection C, however, introduces data auditing techniques that can be used to assess the reliability of secondary or primary data.

This section contains only a brief introduction to the kinds of measures that need to be taken to ensure high quality primary data collection. Readers who are preparing primary data collection are strongly encouraged to refer to Grosh & Muñoz (1996) and Iarossi (2006). These are indispensable references regarding survey data and include treatment of issues surrounding data quality, survey questionnaire design, sampling, and the management of survey fieldwork.

A. Questionnaire Design

As has been discussed in preceding sections, it is important that evaluators know, with as much precision as is practicable, what indicators they need to collect for their study. This should inform the writing of the questionnaire (also referred to as a survey instrument). Collecting data on indicators that are never used in analysis is a waste of time. Furthermore, a needlessly long questionnaire can adversely affect data quality, as interviewers and respondents get fatigued and make mistakes during long interviews.

It is important for the evaluation team to include, or to consult with, people who have experience in questionnaire design. This will help ensure that the responses given during the interviews are as close as possible to the information the researchers want to collect. (See also subsection 1 below on the value of teamwork in developing a questionnaire.)

Questionnaire design can heavily influence data quality. Respondents may sometimes be confused by the phrasing of some questions and inaccurate in their responses even to questions they do understand. People may systematically over- or under-estimate certain things, and recall is often inaccurate.[19] When it is possible to cross-check data against other sources, it may improve accuracy. For instance, survey teams can ask people about travel times to key destinations but also walk or drive the routes and measure the time directly to verify the accuracy.

Furthermore, respondents may give strategic rather than accurate answers, depending on how they perceive the purpose of the survey. Surveys conducted as part of the evaluation of a particular program may be particularly susceptible to this. Respondents may give answers which they believe will positively influence the likelihood of receiving future benefits. Or they may respond as they believe the government, or some other authority or institution, or even just the interviewer, wants to hear (perhaps even to keep themselves out of trouble). This also highlights the importance of clearly explaining to respondents the purpose of the survey and how the information will be used, as part of asking for their permission to be interviewed before going through the questionnaire. (See Iarossi (2006) chapter 5 on survey respondent participation.)

For example, Rao & Ibáñez (2003) found a curious result in an evaluation of the Jamaica Social Investment Fund: subprojects seemed to be poorly aligned with communities' ex-ante preferences, but most beneficiaries nevertheless expressed satisfaction with them ex-post. Perhaps the respondents did not want to appear ungrateful, or perhaps they thought that expressing satisfaction would make it more likely for the community to receive another subproject.

Carefully designing questions to get at key indicators indirectly may sometimes avoid these kinds of strategically inaccurate responses. For instance, in some contexts,

[19] To get a sense of this, write down answers to these two questions for yourself. How many minutes does it take you to commute to work? What was your total income last calendar year? Time your commute over the next five days, and check your tax or income records for last year. How accurate were your responses?

respondents may deliberately understate their income (say, for fear of being caught evading taxes) but be more inclined to accurately describe their spending (often closely correlated with income) or their housing characteristics (an indicator of wealth). Of course, this must be done with caution and a clear understanding of how the desired indicators relate to the indicators actually being collected. After all, asking a different question generates a different indicator, so the researcher needs to have a solid theoretical reason for believing that the new indicator is related to the desired indicator in a particular way. Breaking questions about aggregate amounts into several questions about component amounts is another possible way to get around strategic answers, similar to breaking down questions that are difficult to recall. For example, questionnaires rarely ask "How much did you spend on food last month?" since this is difficult to remember precisely. Instead, they tend to ask about spending on specific food items, and then aggregate the responses to get a more accurate measure of total spending on food. A similar procedure could be used in cases where researchers are concerned that respondents may (deliberately or unintentionally) report false values in aggregate but be more likely to report accurate values for smaller components that make up an aggregate measure.

1. Teamwork

Developing a new questionnaire is a big undertaking, and it benefits from having a variety of people engaged in the task. See Grosh & Muñoz (1996) chapter 3 for an informative presentation on the questionnaire development process. In particular, they point out that analysts, project managers, policymakers, and data managers should all be actively engaged in drafting and refining a survey questionnaire. Pulling together such diverse actors may imply spending more time on questionnaire development than it would take a single tightly-knit team, but the final product will certainly be better and more relevant to all involved. In program evaluation in particular, it is essential that the evaluation team communicate with the program managers so that everyone has a clear idea of the objectives of the program and how these are to be evaluated.

The process of developing the questionnaire is an iterative one, with early drafts being compiled from input from various parties and then subsequent rounds of feedback and revision. Questionnaires typically go through many rounds of revision and editing before a final version is considered good enough and agreed upon. Once the team has finalized the questionnaire to their satisfaction, field-testing it prior to full-scale rollout of the survey is indispensable (see subsection 6 below).

2. Question Wording and Sequencing

The wording and sequencing of questions can have a very strong influence on how survey respondents answer them. Iarossi (2006) chapter 3 has a very informative presentation on question design, which goes into far more detail than is feasible here. Readers are encouraged to consult it before drafting a survey questionnaire, especially if doing so for the first time. He summarizes the key criteria for question wording with the mnemonic "BOSS" for brief, objective, simple, specific.

Brief: Questions should be short. This reduces the chance of misinterpretation through the interviewer misreading the question, or through the respondent misunderstanding it. Brevity also means that each question should only ask one question, with no imbedded assumptions. For instance, "Where do you go for medical services?" assumes that the respondent seeks out medical services in the first place. This may not be true, for any number of reasons. It would be better to break this up into at least three questions. The first asks if the household uses medical services, and then tells the interviewer to skip to the appropriate next question, depending on the response. If they do, the next question asks what type of facility or facilities they have used. If they do not, the next question could ask the primary reason for not using medical services (but only if this indicator is important for the analysis). For data collected specifically for program evaluations, questionnaires should also be kept as short as they can be, collecting only the information that is needed for the analysis. (See subsection VI.E on questionnaire length and sample size.)

Objective: Questions should be as neutral as possible and contain no imbedded suggestions or assumptions as to how the respondent should or will answer. This includes avoiding emotionally charged words, and even the way in which respondents are asked to choose from a list of options. When reading a long list of possible options, respondents may be more likely to choose those toward the beginning of the list. When listening, respondents may be more likely to choose those listed toward the end.[20] The sequencing of questions can also affect objectivity. For instance, questions about engaging in activities or behaviors involving risk might get different responses depending on whether they were preceded by questions intended to gauge awareness of the potential consequences of engaging in risky activities.

Simple: Questions should use the simplest language that clearly communicates the idea. Questions should be written so as to be easily understood by potential respondents with the lowest levels of education. This means using simple, direct words and phrases, and avoiding jargon and technical terms. It also means avoiding negatives and double negatives in phrasing questions. When it is impossible to avoid introducing terminology which must be explained to a respondent, the definition should be given before the question in which the new term appears.[21] Questions which require units of measure should allow the respondent to use the unit of measure with which he is most familiar. The data can later be converted to common units for analysis.

Specific: Questions need to be precise and unambiguous. Words of frequency like often, recently, never, etc. should be avoided. Instead, specific periods the researcher is interested in should be clearly defined. For closed-ended questions, the possible answers should be mutually exclusive and collectively exhaustive. A single question should not have a second question embedded in it. For instance, "How many days did your child go to school last week?" is not, by itself, a good survey question, for two reasons. It implicitly assumes that last week was a normal week in the academic calendar, and "last week" is ambiguous. Last week should be defined clearly either as, say, the Monday

[20] Iarossi (2006), p. 35.
[21] Iarossi (2006), p. 37.

through Sunday preceding today, or as the past seven days. The survey should ask two additional questions: how many days school was supposed to be in session last week and how many days classes were actually held last week.

3. Consistency Checks

Whenever internal consistency checks can be built into a questionnaire, they can save headaches in the later stages of data coding and cleaning. For instance, if asking a farmer about the number of hectares dedicated to each crop, it is wise to add them up on the spot and confirm that they do not exceed the totality of the land he cultivates (accounting for land with mixed crops). Likewise, years of schooling which are within, say, four years of someone's age should trigger a flag. Doing this systematically means incorporating these consistency checks into the layout of the questionnaire, so the interviewer sees clearly that she needs to do them before she proceeds, and in such a way as to facilitate the calculation. For instance, putting column totals at the bottom of each page of an expenditure module makes it clear to the interviewer to calculate the sums before proceeding, and serves as a chance to confirm with the respondents that the aggregate amounts spent in certain categories make sense. Subsection V.C.1 below discusses consistency checks in more detail.

4. Item Non-Response

Item non-response refers to questions which are not answered by a respondent in a survey. (Unit non-response refers to households (or other units of observation) which cannot be found or refuse to participate in the survey, which is discussed briefly in subsection V.B.1 below.) Item non-response codes should be clear, and allow for different reasons for non-response. A combined approach of questionnaire design and coding can reduce non-response confusion.

Careful questionnaire design should make clear which type of non-response is possible for a given question. Questions should be crafted to address only one point without any imbedded questions or assumptions, and skip patterns should be used to make clear why certain questions are not answered. That is, rather than going straight to a question assumed to be relevant to most respondents, it is often better to precede it with two or three brief questions (about existence, availability, and use of a facility or service, for instance) that ensure that the question is relevant and that the respondent is willing to answer it.[22] However, it is difficult in designing a questionnaire to anticipate all the questions and reasons for which non-response could occur.

Additionally, unique codes for various reasons for non-response should be used throughout a questionnaire. That is, "not provided, or not available," "refuse to answer," "don't know," and "not applicable" should all have distinct codes so as not to lose information. The difference between "not provided" and "not applicable" is clearest when thinking about services. For instance, when asking about utilization of health services, "not provided" (i.e., no clinic within reasonable distance) is quite different from "not

[22] See Figure 6.1 in Iarossi (2006), p. 190.

applicable" (i.e., no need to consult medical services, due to good health).[23] These could simply be recorded on a paper questionnaire as non-numeric abbreviations such as DK for "don't know," REF for "refuse to answer," etc., which are used consistently throughout the entire questionnaire.

These can then be converted to unique codes at the time of digitization of the data, preferably using distinct codes for missing data,[24] which are used consistently in the entire dataset. If multiple codes for missing observations are not available in the software being used, another good option is to assign negative numbers which fall outside the range of any possible answer given anywhere in the questionnaire to various non-responses (e.g., -999 for "don't know," -998 for "does not exist," etc.). Negative values are preferable since they are rarely in the range of possible responses to most questions on surveys. This avoids the confusion of cases like the Demographic and Health Surveys' use of 98 as a code for "don't know" when asking the age of household members. Zero should never be used as a non-response code, since many questions may have zero as a legitimate and relevant response, and using zero for anything else would lead to confusion and poor quality data.[25]

5. Copying Former Questionnaires

A couple other issues regarding questionnaire design have been discussed elsewhere in this paper but bear repeating here. It is often a good practice to copy questions from questionnaires used in prior surveys in the country, particularly the census and national household survey. This improves comparability of indicators across data sources and helps ensure that the questions are posed in a way that makes sense in the local context. Of course, if questions in existing surveys are found to be lacking in some respect that is important for the purposes of the new survey, researchers should exercise their professional discretion and make modifications they consider necessary. But they should do so bearing in mind that these modifications imply a cost in lost comparability that should be offset by the benefit of greater clarity or precision. Further, this should not imply limiting oneself to questions in existing surveys when writing a new questionnaire. The motivation for running a new survey is often to ask questions about topics that have not been covered before.

When collecting panel data, the need to adhere strictly to previous survey questionnaires is central. Changing the wording, or even the sequencing, of questions is, in essence, asking a new question. One can no longer treat the indicators as measuring the same thing in different time periods. Again, this does not mean that content cannot be dropped or added relative to prior survey instruments, only that doing so carries costs in terms of comparability and that these should be recognized as decisions are made regarding questionnaire design.

---

[23] Iarossi (2006) chapter 6.

[24] Newer versions of Stata, for instance, allow for up to 27 different codes for missing data, using ".a", ".b", etc., which can be distinguished from each other and from the standard "." that Stata assigns to missing observations. These could be used to distinguish various reasons for non-response.

[25] Iarossi (2006) chapter 6.

6. Field Testing

Even after all of the above steps have been completed and verified in the questionnaire design, the critical final step is piloting the questionnaire with real respondents in the field. It is impossible to predict and eliminate all the possible sources of confusion or errors in the questionnaire in the office. Taking the questionnaire for a test drive will always reveal problems that were not anticipated by the development team.

As with questionnaire development in general, field testing is an iterative process. The team goes through the questionnaire with a few respondents similar to those in the target sample. Problems are noted, and the questionnaire is altered accordingly. The new version of the questionnaire is then field tested, with remaining issues being noted and corrected. The process is repeated until the questionnaire is deemed clear to all respondents, flows clearly and smoothly with all the skip patterns working as intended, and takes an acceptable amount of time to administer.

An additional benefit to piloting the questionnaire is that it also gives the evaluation team the opportunity to pilot test the survey firm and the individual interviewers (also called survey enumerators). If it is clear that the firm contracted to administer the survey is having difficulty doing so satisfactorily, there may be time to re-tender the contract and search for a firm better suited to the task.[26] Similarly, if individual interviewers are having trouble absorbing the training or are found to be shirking their duties in the field,[27] this gives the team the chance to let them go before they can adversely impact the real data collection. It is a common practice to deliberately hire on a provisional basis more interviewers at the pilot stage than will be needed to administer the survey at scale, with the express intention of only making a final selection and only hiring the number needed.[28]

B. Quality Control

Questionnaire design is a crucial component to getting good quality data, but even the most meticulously designed and thoroughly piloted questionnaires can still result in bad data if sufficient quality control is not present throughout the data collection process. Grosh and Muñoz (1996) in chapter 5 offer valuable advice on how to structure and manage survey field operations, and chapters 6 in both Grosh & Muñoz (1996) and Iarossi (2006) contain more details and examples related to data management. This subsection offers a very brief overview of their material.

1. Fieldwork and Data Management

Questionnaires completed in the field may contain errors, ambiguities, or missing information. Checking over questionnaires in the field immediately after they are filled out can reduce the occurrence of these problems. Field supervisors should check forms for completeness, accuracy, and consistency. Checking completeness means making sure

[26] See also Iarossi (2006), pp. 12-18 on Survey Firm Selection.
[27] See also subsection V.C on using Benford's Law to detect when interviewers are fabricating data.
[28] Grosh & Muñoz (1996), p. 106.

that no relevant and applicable questions or sections have been inadvertently skipped. Confirming accuracy means that the answer to each question is clear (i.e., legible and unambiguous). And verifying consistency means checking that embedded consistency checks in the questionnaire have been done and calculated correctly.[29]

The details of structuring field survey teams and the particular duties of each person are too involved to give them adequate treatment here. Grosh & Muñoz (1996) in chapter 5 cover the following topics related to survey field operations:

- composition of survey staff and field survey teams
- specific duties of each team member
- preparation and management of fieldwork
- proper documentation of sample selection and fieldwork procedures

Another key element of good fieldwork is ensuring collaboration by households in the sample. Having households refuse to participate in the survey has negative implications for the sampling design of the survey which adversely affect the rigor of the later analysis using the data.[30] Iarossi (2006) in chapter 5 and Grosh & Muñoz (1996), pp. 121-123, both discuss ways of minimizing household refusal to participate in surveys.

Additionally, survey development and administration should incorporate methods for reducing, or at least flagging, data errors in the field. This can be accomplished using measures such as building consistency checks into questionnaires[31] and direct electronic entry in the field.[32] The key point is to go through as many steps that will reveal errors as possible *while the team is still in the field* and can revisit households for clarifications and corrections if needed.

If there is concern about data entry errors when transcribing information from paper forms into a computer database, it is often worthwhile to have each questionnaire entered twice, by two different individuals, and then compare the two versions for inconsistencies, check them against the questionnaire completed in the field, and retain the correct version.

Proper documentation of all survey design and fieldwork is a key element of quality control when collecting new data and an essential tool for understanding secondary data. Any time a dataset is generated, a parallel process of developing manuals and documentation describing the process by which the data were generated should take place. Manuals should describe clearly how survey fieldwork was carried out, the roles of all survey team members, and include all materials that were used to train the survey teams prior to fielding the survey. They should also describe how field supervision was carried out and include copies of all field supervision and survey administration forms, as well as copies of all questionnaires. Detailed documentation of how the sample was selected and how unit non-response was dealt with is important information that analysts

[29] See subsection V.C.1 below on consistency checks.
[30] See section VI on sampling.
[31] See subsection V.C.1 below.
[32] See subsection V.B.2 below.

need to be able to properly use the data for evaluations or other research.[33] (See also Section VI on Sampling below.)

2. Electronic Field Survey Forms

Instead of using paper and pencil survey forms, electronic data capture in the field can help improve data quality in two ways. One is to alert the survey team to potential errors while they are still in the field (even during the interview) and can revisit respondents for corrections and clarifications. The other is to reduce the scope for interviewer errors like entering non-valid or inconsistent response codes or failing to follow skip patterns correctly. This can be done by giving each enumerator (or each team supervisor) a handheld or laptop computer, where responses can be entered directly, as opposed to recording them on paper forms and later transferring them to electronic media.

Internal consistency checks (discussed below in subsection V.C.1) can be programmed into the software where the data is entered, and alert the enumerator to contradictions in real time, so they can be discussed with the survey respondent and corrected immediately. The data entry software is also programmed to allow only valid responses, although means of overriding this should also be incorporated to allow for non-response to questions, or to signal that additional notes on the response have been recorded on a separate sheet.

If giving all interviewers a handheld computer is not practicable, even having a person on each survey team who enters paper questionnaires into a laptop computer as they are delivered to her by the enumerators in her team can allow discrepancies to be identified, reviewed by the enumerator, discussed again with respondents when necessary, and corrected, *all while the team is still in the field.*

Another benefit of entering data electronically in the field is that it gets the data from field collection to analysis more quickly. Researchers can start analyzing a partial dataset while the survey team is still in the field. This allows them to start testing and refining the hypotheses of the study earlier, and possibly add a question or two to the survey. If the steps outlined above for developing the questionnaire were all carried out, modifying the questionnaire once it has gone into the field for the full-scale survey should only be done if deemed necessary for the success of the evaluation, since these modifications will not undergo the same rigorous vetting process, and since they will not be collected for the full sample.

Analysts can also test the quality of primary data as it is being collected, as soon as it can be compiled electronically in a database, and look for abnormalities by question (possibly indicating inaccurate answers to the question) and by enumerator (possibly indicating fabrication of responses). See subsection V.C below on data auditing techniques to detect potential problems in data.

---

[33] For more on survey documentation, see Grosh & Muñoz (1996), pp. 110-118, and United Nations (2005), p. 32.

3. Checking Secondary Data

For secondary data, many of these same quality control procedures should be replicated when practicable. Observations with contradictory indicators (e.g., years of experience greater than age) must often be dropped, if there is no way to verify the correct information and clean the dataset. Obtaining the data in its raw form, unsorted, as first entered electronically, can sometimes help identify keystroke errors that lead to some inconsistencies. For instance, when finding a household identification code that differs by one digit from those around it, while the sequence of individual identifiers in the next field proceeds sequentially, the researcher may feel confident that this is a data entry error and make the necessary correction. This is obviously not an entirely reliable verification process. Still, it may save a few observations that would otherwise need to be dropped.

Another simple step in ensuring secondary data quality is to make sure you understand how the variables you are using have been coded. For instance, are missing values coded as missing, or have they been entered as something like 99? Is educational attainment entered as years of education, or is categorical (e.g., 1 means some primary, 2 completed primary, 3 some secondary, etc.). It takes little time to check but if overlooked can lead to much time wasted on trying to understand bizarre regression results.

If researchers are in doubt about the veracity of particular indicators in a dataset, they can sometimes use statistical tests to verify whether the data display properties consistent with data from actual observations, as is presented in the following subsection.

C. Data Auditing

If researchers are in doubt about the accuracy of particular indicators in a dataset, they can use a variety of techniques to verify whether the data display properties consistent with data from actual observations. These methods can be used to detect errors in recording responses on the questionnaire, errors in data entry, poor understanding of questions by survey respondents, and fabrication of data.

This section describes some of the checks that can be used to assess the quality of datasets, both primary and secondary. Of course, when it is possible to integrate computer-based quality controls to fieldwork, many of these tests could and should be incorporated to the survey's data entry program or be made a part of the observations made by the survey project managers while the survey is being fielded. The checks can then enable the implementation of effective corrective actions and improve the quality of data collected in the field, as has been discussed earlier in this section. The following discussion highlights a few particular checks that can be incorporated to reduce errors.

1. Internal Consistency of Individual Observations

Almost every dataset has some degree of internal redundancy that can be used advantageously to assess its quality. This subsection presents some common examples which should help the reader come up with appropriate checks for her own data.

Internally inconsistent responses (e.g., a difference of less than four between age and years of schooling) can be used to flag potential errors in data collection for particular observations. But a high proportion of observations with internal inconsistencies may indicate a flaw in the data collection instrument, rather than errors in the data collection process. Hence running these checks on data collected while pilot testing a survey instrument can be useful for finding problems with a questionnaire that might not be apparent otherwise.

Monetary Balances: The total revenues and total expenses of a household or of an enterprise cannot be expected to match exactly—not even theoretically—in a given period of time. However, large discrepancies between the two figures for a large proportion of observations can reveal shortcomings of the data collection process, if complete measurement of income and expenditures is intended. This is often associated with the omission of specific budgetary items—such as income from self-employment, pensions, or remittances—or failing to account for categories of expenditures or savings. This check can also reveal fabrication, if too many exact matches are found.

Food Intake vs. Household Composition: The reduction of malnutrition is an objective of many development projects, but even in the broader context of poverty reduction, the recording of food consumption is often a central element of a household survey expected to evaluate a program's impact. Unfortunately, accurate measurement of food consumption is very difficult. It is thus important to compare the energy intake implied by the recorded food consumption with the composition of the household. This need not require deep knowledge of nutrition—a rough and simple calculation of each household's energy intake per capita (not even per adult equivalent) will be sufficient to reveal such shortcomings if they exist. A researcher might be concerned with the accuracy of the data if he finds too many households in the dataset which appear to be consuming less than, say, 800 or more than 5,000 kilocalories per person each day. Observations with such values might be flagged as possible errors, to be verified in a second visit to the household. If computers are used to record responses directly during the interview, an internal check can be programmed into the software, which asks the interviewer to double-check answers with the respondent if the values fall too far outside normal ranges. However, finding some households in these conditions is normal and can be explained by special circumstances. Households may appear to be consuming too few calories if members have meals outside home, or to be consuming too many if they happen to be interviewed during a festival and are hosting a lot of guests.

Consistency of Anthropometric Measures: The consistency of height, weight and age measures can be checked against the Word Health Organization reference tables.[34] These tables can be used to compute Z-scores[35] for children's Stature-for-Age, Weight-

[34] http://www.who.int/childgrowth/standards/technical_report/en/index.html. The use of the WHO tables rather than other standards can be a matter of endless debate. This typically goes into a level of refinement which is beyond that necessary for data auditing, which merely requires rough measures in order to detect large deviations from expected distributions.

[35] The Z-score of a variable X is defined as $Z_i = (X_i - \mu_X) / \sigma_X$, where μ_X and σ_X are the mean and the standard deviation of X.

for-Age, and Weight-for-Stature, and for adults' Body Mass Index. A high prevalence of Z-scores with absolute values greater than 4 may indicate that the measures are unreliable.

Consistency between the Cost and Quantity of Purchases: This check should be implemented in all surveys that record both costs and quantities. It is in fact one of the main reasons for recording quantities in surveys where costs are arguably sufficient. It requires establishing acceptable boundaries for unit prices. These boundaries may obviously depend on the product (a kilo of tomatoes is cheaper than a kilo of meat), and on season and location (tomatoes are cheaper in the summertime in the countryside). For the purposes of data auditing, however, the later can often be ignored, and very broad item-specific ranges are sufficient (for instance, one kilo of tomatoes should cost between 3 and 10 pesos, regardless of location and season).

Consistency between Expenditures and Other Characteristics of the Household or Enterprise: This is a very broad category of quality controls based on the idea that expenditures on certain items—such as electricity or diapers—should occur only where consistent with other characteristics—such as being connected to the electric grid or having babies present in the household.

Using Built-In Redundancies: Survey instruments often contain redundancies deliberately included with a view to quality control. These may include control totals (for instance, a row for total expenditure on food at the bottom of a page with detailed expenditures on specific food items), inquiring about both age and birth date, etc. When such features are available, they should obviously be used for auditing.

2. Consistency with External Data Sources

It is sometimes possible to compare a survey's estimates with figures from alternative sources. Matches are almost impossible—even within the limits of sampling error—but if the orders of magnitude are too different, this probably indicates problems with one of the data sources. Some examples relevant to impact evaluation are:

- comparing the regional, age and gender distributions of a population, as estimated from the survey, with the corresponding demographic projections, based on the latest census
- comparing school enrollment and the utilization of health services with corresponding figures from the ministries of education and health
- comparing the number of beneficiaries (or benefits such as water pumps) as estimated from the survey, with the number of beneficiaries (or benefits), as reported from a project's administrative sources. The fact that these figures almost never match is one of the main reasons for inquiring about benefits in household surveys, but extreme differences could reveal shortcomings in survey fieldwork rather than simple leakage.

3. Consistency among Subsets of a Dataset

Surveys are sometimes fielded over a period of several months, or even over the course of a full year. Under these conditions it may be useful to compare the behavior of certain indicators over time, looking for patterns that may reveal shortcomings of supervision rather than seasonal or secular trends. When such patterns are detected, they can generally be associated with specific fieldworkers or data collection teams. Examples include:

- reduction in the average household size, or in the number of persons in specific gender/age groups. This can happen when unsupervised interviewers realize that they have to work harder in larger households than in smaller households.
- reduction in the number of transactions recorded in each household
- reduction in the recording of certain events (such as episodes of illness or injury) or activities (such as self-employment), in the absence of which whole sections or modules of the questionnaire can be skipped

It may occasionally be possible to detect suspicious differences among fieldworkers even in surveys that are not fielded over long periods. The so called ***interpenetrating samples*** (Mahalanobis, 1946) are deliberately designed to have different interviewers work in comparable areas, so that any significant differences among interviewers can only be attributed to fieldwork management (or mismanagement.)

4. Statistical Checks

Careless or fraudulent reporting may sometimes be detected by observing the statistical distribution of the *last* digit of some indicators:

- The tendency of respondents and interviewers to report ages ending in zero or five years (or in multiples of twelve months in the case of young children) is a well known demographic problem, which can be reduced with appropriate instructions to the interviewers, and with survey instruments such as a calendar of events. Tabulating the distribution of the last digit can thus reveal deficient training or supervision.
- The use of rounded figures in the recording of height or weight does not depend on the respondent because those figures are the result of instrumental measurements. Therefore a high incidence of values for height ending in zero or five millimeters can only be attributed to careless interviewer behavior. The same is true for the weights, which are generally measured in kilograms with one decimal.

Benford's Law[37] is a statistical phenomenon which many survey data adhere to and which can be used to detect abnormalities in some survey data. For evaluators who want to check the quality of secondary data they might wish to use, engaging a statistician to compare indicators of interest from the existing data against Benford's Law is a good first step toward verifying that the data do not suffer from any obvious problems. This technique can be used to detect if data is being fabricated by survey enumerators or firms, and flag questions which respondents find difficult to understand or recall.

Table 1: Distribution of First Significant Digits according to Benford's Law [36]

Leading Digit	Probability
1	30.1 %
2	17.6 %
3	12.5 %
4	9.7 %
5	7.9 %
6	6.7 %
7	5.8 %
8	5.1 %
9	4.6 %

Judge and Schechter (2006) demonstrate how the application of Benford's Law was used to detect errors in socioeconomic household survey data. They show how this technique can be used to detect enumerator falsification of survey data, and errors of comprehension or recall on the part of respondents. John Morrow, a colleague of Schechter's, has developed an online tool where anyone can check their data to see if it is consistent with the distribution suggested by Benford's Law, which can be found at http://checkyourdata.com/.

An exposition of the mathematical properties of Benford's Law is beyond the scope of this paper. But in essence it states that, in data in which the spread of values covers at least an order of magnitude, the first significant digits are uniformly distributed on a log scale. This means that, for indicators which have a highest value that is at least ten times the lowest value, each digit appears as the first non-zero digit in a number with frequencies close to those displayed in Table 1. This occurs regardless of units of measurement. However, data that do not take on values over a range of at least one order of magnitude, such as adult height or weight, will not follow Benford's Law.

VI. Sampling

If an evaluation involves primary data collection, a sample must be constructed. Several aspects must be considered:

- sampling frames for each stage of sampling
- stratification, oversampling, and sample weights
- sample size and statistical power
- selection of comparison groups
- how new data might build on existing data to construct a panel dataset
- trade-offs between questionnaire length and sample size
- documentation of sample selection procedure and field implementation

[36] Source: Judge & Schechter (2006)

[37] Those who want a more thorough introduction to Benford's Law than is within the scope of this paper can browse http://www.google.com/search?hl=en&q=benford%27s+law&btnG=Google+Search.

These points are largely interrelated and must be managed simultaneously, as decisions related to one aspect influence others.

Iarossi (2006) chapter 4, United Nations (2005) chapter 2, and Grosh & Muñoz (1996) chapter 4 all have detailed discussions of sampling, and readers looking for guidance in constructing a sample for a survey are encouraged to refer to those sources in addition to this section. Constructing a sample for data collection is a complicated business. Many different factors must be considered, best guesses made about some parameters, and judgment calls made about trade-offs of various sampling schemes. The process is mathematical in essence, but it involves nuanced decisions that depend on the objectives of the data collection and the nature of the study. If an evaluation team decides to collect primary data, having a team member or consulting with someone who has experience in sample selection for surveys is advisable.

A. Multi-Stage Sampling and Sampling Frames

Some surveys collected or used for evaluations may employ a single-stage sampling procedure, such as selecting individual applicants to a program from an administrative list of all applicants (whether they participated or not) or selecting registered companies from a national list of firms maintained by a chamber of commerce or government entity. But since household surveys are the most common source of data used in impact evaluations—and since household survey samples are usually selected using a two-stage sampling procedure—this section will focus on two-stage sampling. The mechanics of one-stage or multi-stage sampling are similar to the relevant steps for two-stage sampling.

1. Two-Stage Sampling

For household surveys, a random draw from the full population of potential observations is rare in practice. Most surveys use two-stage sampling as a pragmatic way to concentrate field activities in order to simplify the logistics of data collection fieldwork.

The first stage is random selection of primary sampling units (PSUs). PSUs are typically small geographical units with 50-200 households, and are most often selected from a list of census enumeration areas (also called census tracts) maintained by the national statistical agency from the most recent census. (See subsection 2 below on sample frames.) Giving each PSU a probability of being selected in proportion to its size (number of households, or population) gives all potential observations an equal chance of being selected into the sample, which ensures that the sample is representative of the population of interest. Grosh & Muñoz (1996), pp. 66-79, present a straightforward method for selecting PSUs with probability proportional to size. United Nations (2005), pp. 319-332, also presents an example of multi-stage sampling with probability proportional to size using Excel.

The second stage is random selection of units of observation (most often dwelling units as a proxy for households) within each selected PSU. This second stage usually requires the creation of a new listing of all the potential units of observation in the selected PSUs. In the case of a household survey, this means visiting each PSU to map out and list all dwellings, so a sample can be selected from an exhaustive list of potential observations.

In two-stage sampling, the number of PSUs multiplied by the number of observations per PSU equals the total sample size. But determining the number of observations per PSU is at the discretion of the evaluation team, and it influences the total sample size needed for a given level of statistical power. (See subsection VI.C below on how sample size is determined and how multi-stage sampling must be accounted for in sample size calculations with cluster effects.) This determination must take account of the nature of the intervention and the hypotheses the evaluation intends to test. For instance, if analysis of impacts across individuals within communities is of interest, it will be necessary to have more observations per PSU for a sample to be representative at the village level. However, if the thrust of the research is more in terms of measuring impacts across communities, the sample should have more PSUs and relatively few observations per PSU, to account for clustering of effects at the community level.

Furthermore, determining the number of observations per PSU must also take into consideration the logistics of running the survey in the field. For this reason, observations per PSU are typically some multiple of the number of interviewers in each survey team. For instance, if each survey team consists of four interviewers and one supervisor, the number of households to interview per PSU would be some multiple of four. This way, since the four interviewers would be working simultaneously, the team members would all be ready to travel to the next PSU at roughly the same time. The average time needed to get through the questionnaire and the travel time between PSUs also influence the logistics of fieldwork and should be considered as well.

2. Sample Frames

A sampling frame (or sample frame) is a list from which a sample can be selected, and one is needed for each stage of sampling. As such, it must include (or be representative of) the entire population of interest for the evaluation. Sometimes, a valid sampling frame already exists from which a sample can be drawn. For instance, as mentioned above, a common practice is to use the list of census tracts from the most recent population and housing census to select PSUs.

If the census is less than a year old and can be accessed in sufficient detail to identify and revisit selected households, it could be used to select observations within each PSU as well. However, since people move around, a second-stage sample frame must be recently constructed, or it will bias the sample toward those who stay in the same place over time. Often, a ready-made second-stage sampling frame is not available, and the evaluation team must carry out a complete listing of households (or more often in practice, residence units) in each of the selected PSUs.

Care needs to be taken in the selection method and size of the sample, and these considerations will influence how the sample frames are constructed. Samples may want to include only observations that meet certain criteria (e.g., households with children under five years of age), which clearly implies that the second-stage sample frame should include only units targeted to be in the sample. These issues are discussed in subsection VI.B on sample stratification and related issues and in subsection VI.C on sample size and statistical power calculations below.

B. Stratification, Oversampling, and Sample Weights

Sample stratification, *oversampling*, and *sample weights* are related concepts that often come into play when constructing a sample for primary data collection. Because many national household surveys use stratified samples (e.g., to capture representative samples of urban and rural populations) and some oversample some areas or groups, sample weights are a common field included in these datasets, so knowing how to incorporate them into analysis is useful for using secondary data.

1. Sample Stratification

Stratification refers to partitioning the population of interest into mutually exclusive and collectively exhaustive subgroups, from which samples are then drawn. It is used for two purposes. One is to improve statistical power for a given sample size, by sampling more observations from strata with higher variance along a variable of interest (e.g., sampling more households in higher income brackets if income is a variable of interest). (See United Nations (2005), pp. 19-20; Grosh & Muñoz (1996), pp. 58-59; and subsection VI.C below on how design effects need to be incorporated in sample size calculations.)

The second purpose for stratification is to ensure a sufficient number of observations from subgroups of interest (referred to as analytical domains).[38] It is particularly useful to ensure that minority groups within the population are adequately represented in the sample to allow for statistical inference. Take for example a country with four linguistic groups—A, B, C, D—making up 50%, 25%, 15%, and 10% of the total population, respectively. A small sample of 300 households might be stratified accordingly (150 of A, 75 of B, 45 of C, 30 of D) to ensure each group is adequately represented, since a random sample might end up with too few observations from group D or even group C. (This is assuming that the evaluation wants to be able to differentiate impacts by linguistic groups, or use mother language as a control variable.) Alternatively, if the evaluation is interested in differentiating effects within linguistic groups, the sample might be constructed by selecting 75 observations from each group. Drawing inferences on the overall population would then require the use of sample weights in the analysis to account for the fact that group A is underrepresented in the sample, while groups C and D are overrepresented. (See subsections 2 and 3 below on oversampling and sample weights.)

[38] Some authors say that, strictly speaking, stratification refers only to partitioning the population to improve statistical power. Since the procedure is identical for both purposes, this paper will refer to it as stratification, regardless of the underlying motivation.

Stratification by groups within the population requires a sample frame (see subsection VI.A.2 above) that identifies observations as belonging to a particular group. In the above example, one could use a recent census if it included a question about which language is spoken at home, so that census tracts could be stratified by the predominant language spoken in the area.

2. Oversampling

If an evaluation design calls for analysis of project impacts on different groups within the overall population of beneficiaries, the sampling strategy must sometimes adjust to allow for this, by oversampling groups that are not very prevalent in the wider population.

Say, for instance, that the effects of a particular program are suspected to be different for young children of illiterate mothers than for young children whose mothers can read. In a country with widespread female illiteracy, this may not alter the sampling strategy, since a sample which is representative of the overall population will contain many observations both of households with illiterate mothers and of households with literate mothers. If, however, the illiteracy rate for women in the relevant age range is fairly low, it may be necessary to oversample this group in order to have a sufficient number of observations for meaningful statistical analysis (see subsection VI.C below on sample size and statistical power). This means constructing the sample so that it includes a higher proportion of households with illiterate mothers than is found in the overall population.

Oversampling is done by stratifying the sample, and assigning a higher number of observations to an oversampled stratum than it would get under a proportional sample. To continue the example, using a recent census that includes a question on whether the mother can read and write would suffice to oversample households with illiterate mothers, since census tracts with high proportions of illiterate mothers could be identified.

Sometimes, the evaluation team may have a sample frame that does not include an indicator that allows the identification of the group they want to oversample. They must then decide whether oversampling this group is worth the additional expense and complication of augmenting the sample frame to allow identification of groups of interest, which could imply constructing a new sample frame for one of the stages of sampling. Say, for instance, that the evaluation team had a five-year-old census and wanted the sample to include a minimum number of illiterate mothers with children under five. The second-stage sample frames (the household listings for each PSU) would then have to include information on female literacy and age of children for each dwelling listed.

3. Sample Weights

Sample weights (or sampling weights) are used to account for the fact that a sample has been constructed in such a way that each potential observation did not have

an equal probability of being selected. The sampling weight for each observation is equal to (or proportional to) the inverse of its probability of having been selected. So each observation's weight is (proportional to) the number of potential observations in the category from which it was selected, divided by the number of observations from that category in the sample. (See the last paragraph of this subsection for an example with numbers.) Sampling weights need to be calculated and included in all primary datasets (except in cases of self-weighting samples, described in the next paragraph). And they need to be used in analysis of secondary datasets for statistical inference to be unbiased and accurate.

Samples that give equal probability of selection to all potential observations in the overall population, such as a one-stage random selection from the full population of interest, are called self-weighted (or self-weighting). That is, each observation has the same sample weight. Since each potential observation had an equal probability of being selected into the sample, there is no need to adjust the sample weights. Since the ratios of sample weights matter, and the absolute values do not, it is simplest to let the weights equal one in a self-weighted sample.

However, when multi-stage sampling, stratification, or oversampling is used, sample weights are used to account for the higher probability of selecting certain groups. When a sample is not selected with equal probabilities on all potential observations, observations must be weighted to produce statistics regarding the population at large. This means that oversampled groups are assigned weights of less than one, and undersampled groups weights greater than one, such that the weighted sample has proportions which accurately correspond to the overall population. This is a common practice in national household surveys, where urban areas may be oversampled to account for higher variation, and more homogenous rural areas may be undersampled. Sample weights are then assigned to each observation, proportional to the number of individuals in the overall population that the observation represents.

Even in cases where multi-stage sampling is designed to be self-weighted, by selecting PSUs with probability in proportion to size, calculating and including sample weights may be necessary. In the case of a two-stage household survey sample, the listing of households in the selected PSUs may show that the number of households has changed in some PSUs disproportionately to others. For instance, some PSUs could have more households than was initially believed, while others have gained none, or even lost some. Or all PSUs could have grown, but with some having doubled their number of households, whereas others grew by only ten percent. If the growth rate was not the same across all PSUs, sample weights must be used.

For a basic example of the calculation of sample weights, say an evaluation team has determined that a sample size of 1000 households is needed, from a population where five percent of households with children age 0-5 have mothers who are illiterate. The team has also determined that the sample should contain at least 100 households with mothers who cannot read. After eliminating households with no children in the appropriate age range, the sample is stratified by mother's literacy. One hundred households are selected from the group with illiterate mothers, and 900 from those with

literate mothers. The sample weights are then 0.5 for illiterate-mother observations (from 5% in the population divided by 10% in the sample), and 1.056 for literate-mother households (from 95% in the population divided by 90% in the sample).

C. Determining Sample Size and Power

It is important to make sure the sample size being used provides sufficient statistical power to detect an effect of interest. This is true whether using primary or secondary data. An "effect of interest" is what policy makers would consider a meaningful change, one that is big enough to justify the cost of the program being evaluated. Assigning statistical significance to an economically insignificant difference is not needed and could waste resources. Throughout this section, it is important to bear in mind that for studies that are looking at multiple outcomes or impacts (i.e., most impact evaluations), sample size calculations should be carried out for each outcome or impact to be studied, since each will potentially yield different numbers. The unconstrained optimal sample size is then the largest necessary, from the calculations for each outcome and impact of interest. In practice, sample sizes are usually limited by budget constraints or for other reasons, and the various sample size results can give the researcher an initial idea regarding which topics are tractable in the evaluation and which may be difficult to get a handle on.

With secondary data, for each outcome of interest, one can calculate the effect size that can be detected with the given sample size and observed standard deviations, for selected levels of statistical confidence and power. If the minimum detectable effect size is much greater than the smallest effect which would be considered economically significant, then analysis showing no impacts could occur for one of two reasons. ("No impacts" refers to coefficient estimates on parameters of interest that are not statistically significantly different from zero, in the language of regression analysis.) One is that the effect really is zero. If the confidence interval around a coefficient estimate contains no values that researchers consider economically significant, the impact probably really is zero. On the other hand, it is possible that the sample size is too small to detect the effect, even though it is not zero. If the confidence interval around a coefficient estimate contains both zero and values considered economically significant, the regression is underpowered, in the language of statistics.

If the evaluation includes the collection of primary data, calculating an adequate sample size for an acceptable level of statistical power is an important honest assessment of whether the resources available to conduct an evaluation are sufficient to provide reliable answers to the questions the evaluator wants to address. Using too small a sample may have a low probability of detecting an important effect that actually happened. Such inconclusive research is of little use or could even be interpreted (wrongly) as "evidence" that the program did not have any effect. The same resources might be put to better use in an alternative research design. Conversely, collecting too large a sample is often not the best use of the limited resources available for an evaluation.

Most impact evaluations look at the effect of a program on more than one outcome. In such cases, sample size calculations should be done for each outcome. If the

largest sample size from this set of calculations is feasible, then it should be used. If, however, the largest feasible sample size falls somewhere in the middle of the set of sample size calculations, the research team may have to consider some trade-offs regarding the scope of topics the evaluation can address and the sample size used. Finally, if the calculations for every outcome yield sample sizes that are larger than what the evaluation budget can afford, the evaluation team may need to think about an alternative research design that is more likely to produce useful results. (Bamberger [2006] offers some suggestions along these lines.)

Subsection 1 below describes the intuition behind determining sample size and statistical power, and subsection 2 then provides some formulas that can be used in relatively simple cases. Subsection 3 then briefly discusses methods for determining sample size when the research framework is too complex for a formula to suffice. Readers who need more detailed information can consult Iarossi (2006) chapter 4, United Nations (2005) pp. 25-29, Cochran (1977), or Lohr (1999).

1. Sample Size Intuition

Determining an appropriate sample size for a study is rarely a simple task, especially if the model being tested is complex. It is often wise to consult with a statistician to confirm that one is going about it properly. But an evaluator can do it himself for a simple study design and can prepare himself for a fruitful conversation with the statistician in any case.

In addition to the sources already cited, three six-page articles by Lenth (2001) and by Eng (2003, 2004) offer intuitive introductions to calculating sample size and statistical power, including formulas for simple analytical models, and discussion of how to use simulation where the standard formulas do not apply, as well as common pitfalls to avoid. A summary of their key points follows here.

The researcher essentially needs five pieces of information to be able to calculate the needed sample size:

1. effect size of interest
2. standard deviation of outcome metric
3. confidence level
4. power level
5. design effect, or cluster effect, of the sample selection procedure

These are explained here. The first two rely on the researcher's expertise in the question to be studied. The third and fourth pertain to the sensitivity of the statistical model. The fifth is determined by how the sample is constructed. A generic example is used to illustrate: Imagine a country with a school year of 150 days, where the average primary student attends 100 days. Students often report illness as the reason for their absence, and malaria is endemic. It is therefore believed that providing bednets for children will increase school attendance by reducing malaria transmission. Children are randomly selected to receive bednets, regardless of current enrollment status.

The first step in determining sample size is determining the effect size one wants to detect. This hinges on what policy makers hope to achieve through the intervention that is being evaluated. Impact evaluations typically involve comparing mean outcomes between treatment and comparison groups. Smaller differences (i.e., smaller effects) need larger samples to detect them. At some point differences are so small as to be economically negligible. One way to avoid unnecessarily large (and expensive) samples is to determine, through consultation with policy makers, the smallest meaningful effect size for a project to be considered a success. The larger an effect size one is willing to "settle for" detecting, the smaller a sample can be. Let us say in our example that, for the cost of a bednet, policy makers consider that an increase in attendance of less than 10 days is not worthwhile. This is our minimum effect size of interest.

The second step for calculating sample size is estimating the standard deviation of the outcome metric over the population of interest. Sometimes existing data can be used to calculate this; sometimes existing literature or previous evaluations on similar topics suggest a consistent standard deviation; sometimes the researcher simply does not know. In the last case, a rule of thumb is to use one quarter of the range as standard deviation. (Since two standard deviations above and below the mean contain about 95% of the observed values in a normal distribution, one quarter of the difference of the highest value minus the lowest value gives a rough approximation of the standard deviation.) In our example, let us say that we have no data which could be used to calculate the standard deviation for primary attendance. But we know that some kids never go to school, and some show up every day. Therefore our range is 150 (from the difference between 150 and 0), and one quarter of that is 37.5, our best guess of the standard deviation for our outcome of interest.

The third step is deciding on a confidence level for our evaluation results. This means choosing the probability (often called "alpha," or "the level of the test") of getting a false positive (also called a "type I error"), that is, the likelihood of finding a seemingly significant effect when the true effect is zero. Obviously, we want this to be low, since attributing results to a truly useless intervention could lead to wasting a lot more money. Common thresholds for alpha are 0.1, 0.05, or 0.01. Samples sizes increase for lower alphas (lower chances of false positive findings). In our example, let us say that we are only willing to accept a 5% chance of a false positive, so our alpha is 0.05. That is, if we detect that children who were given bednets attend school at least ten more days a year than children not given bednets, we want to be at least 95% confident that this difference is systematic and not due to random variation.

The fourth step is deciding the power of the test. This means choosing the probability (often called "beta") of finding a false negative (also called making a "type II error"). The power of the test equals 1-β, that is, the probability of detecting a specified effect if it truly exists. So the power of the test is the probability of avoiding a false negative (or the probability of avoiding a type II error). Obviously, we want power to be high, so we want beta to be low. If the program does have an effect, we want our statistical test to have sufficient power to be able to detect it. Conventionally, a maximum beta of 0.20 is considered acceptable. Sample size increases for lower betas (higher chance of detecting a given effect if it exists). In our example, let us say that we want to

be reasonably sure (90% sure) of detecting a ten-day increase in school attendance if providing bednets actually increases average attendance by at least ten days. So we set our beta to 0.10, to have statistical power of 0.90.

The fifth step is accounting for cluster effects, or other design effects, that stem from how the sample is selected. For two-stage sampling, intra-cluster correlation within PSUs (that is, the degree to which households near each other tend to be alike) reduces statistical power for a given sample size. United Nations (2005), pp. 19-21, and Grosh & Muñoz (1996), pp. 58-59, discuss design effects and illustrate how to adjust for them in sample size calculations. The composition of the two-stage sample affects its total size. A sample with more PSUs and fewer observations in each PSU yields more power but is logistically more difficult and more expensive to manage in the field. A sample with fewer PSUs and more observations per PSU simplifies logistics and reduces the cost of fieldwork, but at the expense of statistical power. Determining the right balance between sample size and composition is a judgment call the evaluation managers and data collection managers must come to a consensus on. Let us assume for the example that attendance records for several schools allow us to estimate that the intra-cluster correlation is 0.02. That is, children are slightly more likely to attend class if more of their classmates in the same school are attending class that day.

Once these five pieces are in place, the sample size can be calculated using formulas presented in Iarossi (2006) chapter 4, or in the next section for a simple study design, or it can be calculated using simulations for more complex study designs (as discussed in subsection VI.C.3 below).

But bear in mind that the calculation yields the sample size needed for use in the statistical analysis. If the evaluator expects some observations to be dropped, due to problems in the field or in processing the data, the sample size needs to be inflated by the expected rate of "lost" observations. In practice, a few observations are inevitably dropped before the data can be analyzed, so the evaluator must anticipate this. For example, in the case of a household survey, if the evaluation team expects five percent of households to be away from home or to refuse to participate, the list for the intended sample needs to be inflated such that a five percent attrition rate still yields the minimum sample size needed. However, sample attrition presents a more serious problem than just a reduction in sample size. It introduces a source of potential bias into the sample, and into any subsequent analysis using the data, that cannot be estimated or controlled for. This is why managing fieldwork to minimize lost observations is so important.

There is no generic panacea for this situation, and the researcher must deal with it as best he can. For instance, if there is a lot of migration during certain months for seasonal agricultural labor, it would best to avoid running a survey that hopes to find agricultural laborers at their permanent residence during these months. Some household surveys routinely include replacement households in the sample selection procedure, which can be used if households in the initial sample cannot be tracked down, or are otherwise impossible to interview. However, allowing for replacement observations still does not eliminate the potential bias caused by failing to capture observations with unknown characteristics. Hence, some researchers prefer not to replace lost observations

at all, lest survey teams be tempted to give up too quickly on including households in the sample which are difficult to find or convince to participate. Pilot studies, discussed in subsection 3 below, can help give a sense for how many observations will likely be dropped during the data collection and perhaps help the research team understand why and how to minimize this number.

2. Sample Size Formulas

In practice, a simple comparison of means in a randomized treatment context is not a common program evaluation design. Typically, evaluations are much more complex, and the evaluator will need to use more complex formulas or even simulations to determine an appropriate sample size.[39] The presentation of the formula here is intended to illustrate the mechanics of how the elements discussed in the previous section come together to determine sample size.

Assuming the sample is divided equally between treatment and control observations, and assuming that both groups have the same standard deviation, the formula is

$$N = \left[\frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[1 + \rho(H-1) \right]$$

where N is the total sample size; D is the difference in means, or effect size, that the evaluator wants to detect; σ is the standard deviation of the outcome metric; $z_{\alpha/2}$ is taken from Table 2 below depending on the selected alpha[40]; z_{β} is taken from Table 3 below depending on the selected beta; ρ is the intra-cluster correlation coefficient; and H is the number of observations sampled in each cluster. As the reader can see, N decreases as D increases, so detecting a larger effect can be done with a smaller sample, and detecting small effects requires a larger sample. Also, a larger standard deviation in the outcome σ means that N is larger for a fixed D. That is, the more the outcome metric varies in the population, the larger a sample is needed to detect an average difference between two subgroups (e.g., treatment and comparison) within that population. Intra-cluster correlation also inflates the necessary sample size. Note that if there is no intra-cluster correlation, ρ is zero, and the second term simplifies to one. However, the higher ρ is, the more observations are needed. Moreover, if ρ is positive, the more observations are clustered in the sample (i.e., the higher H is), the greater the overall sample size needed. This is because additional observations within clusters yield less information, since their intra-cluster "neighbors" are so similar. So a lower H decreases the overall sample size, but at the expense of sampling more clusters, since the number of clusters in the sample is N divided by H.

[39] See Kraemer and Thiemann (1987) for formulas for slightly more complex statistical models.
[40] Alpha is divided by two for a two-tailed test.

Table 2: Setting Confidence Level

Critical values for confidence interval for two-tailed test of difference of means for a normally distributed variable

Confidence Level (α)	$z_{\alpha/2}$
.01 (99%)	2.576
.05 (95%)	1.960
.10 (90%)	1.645

Numbers in parentheses are the probabilities that the confidence interval of the estimated difference of means contains the true difference of means.

Values not shown in this table may be calculated in Excel by using the formula $z_{\alpha/2}$ = NORMSINV(1-$\alpha/2$) where alpha is the accepted probability of finding a false positive.

Table 3: Selecting Statistical Power

Critical values for selected statistical power in a two-tailed test of difference of means for a normally distributed variable

Statistical Power (β)	z_{β}
.80	0.842
.90	1.282
.95	1.645

Power is the probability of avoiding a false negative.

Values not shown in this table can be calculated in Excel by using the formula z_{β} = NORMSINV(β). For calculating power, the inverse formula is β = NORMSDIST(z_{β}), where z_{β} is calculated from the equation in the text by solving for z_{β}.

Source: Eng (2003)

Returning to the example of bednets and school attendance from the previous subsection, we are now ready to determine our necessary sample size using the formula. We set D equal to 10, since any reduction of absenteeism of less than ten days is not considered worth the expense of the program; σ, the standard deviation of attendance, we estimated at 37.5; $z_{\alpha/2}$ is 1.960 for 95% confidence, from Table 1 above; and z_{β} is 1.282 for 90% power, from Table 2 above. The estimate of intra-cluster correlation from previous attendance records from several schools is ρ=0.02. For logistical reasons, the evaluation team decides to sample 20 children in each cluster (each school catchment area). Plugging into the formula, we get the following expression, which yields an N of 815.88, or a sample size of 816 children (408 with bednets and 408 without).

$$N = \left[\frac{4(37.5)^2(1.96+1.282)^2}{(10)^2}\right]\left[1+(0.02)(20-1)\right] = 815.88$$

Stata includes a command, *sampsi*, which can perform this calculation for a simple study design (namely a straightforward comparison of means, as discussed here). However, as of Release 9, Stata does not have an option to account for intra-cluster correlation, so this must be done manually after calculating the sample size under the assumption of no intra-cluster correlation. For this example, the Stata command "sampsi 0 10, sd(37.5) alpha(0.05) power(0.90)" yields a sample size 296 in each group, for a total of 592. This must then be inflated by multiplying it by [1+(0.02)*(20-1)], which yields 816.

But bear in mind that this is the final sample size needed *for analysis*, so if we expect records on some children to be impossible to locate, our sample frame needs to be inflated to compensate for the expected rate of "lost" observations. However, a more serious concern is that these lost observations will bias the sample and the analysis, no matter what the sample size. This is best dealt with by tracking down as many of the lost

observations as possible, to collect the full sample as it was selected. (See Annan, Blattman, and Horton (2006) on tracking down and accounting for former child soldiers for a survey sample in Uganda.)

Let us imagine that we know from experience in this country that attendance records are typically unavailable for about 15% of children, either because the children move during the school year, or because the records are simply lost. Then our sample frame (the list of children for whom we want to collect data) needs to be 18% larger than our final sample size. (The percentages do not match, since they are calculated from different bases. A simple way to calculate the size of the sample frame is to divide the sample size needed for analysis by the percent expected to remain after attrition.) In our example, that means we need an initial list of 960 children (816 divided by 85%) for whom we want to collect attendance records. This gets the sample size right, but it does not correct for the bias inherent in losing observations.

It is worth repeating that there must be no conceivable systematic reason for these lost observations, in order for the sample to be considered unbiased. If richer families are more likely to move, or if some teachers or schools throw out the records for their worst students, the remaining sample will be biased, and this will compromise any subsequent evaluation using this data. The evaluator either needs to find a work-around to construct a truly representative sample, or deliberately restrict the sample to the population it can accurately represent. In the latter case, the analysis and reporting of results need to be very clear about the nature of the sample, and any limitations on generalizing the findings.

The manufactured example we worked through here is plausible but simple. In practice, impact evaluations of development projects are often more complex. For instance, the above discussion ignores the role that covariates, or control variables, play. In multivariate regressions, the inclusion of control variables can increase or reduce statistical power, depending on their correlation with other variables of interest. Inclusion of highly correlated controls reduces statistical power, while inclusion of uncorrelated (or orthogonal) controls increases power. Determining the influence of covariates on the necessary sample size typically requires pilot studies or simulations. Where pilot studies or simulations are not practicable, using formulas like those presented in this subsection and in the references is a decent method for estimating sample size.

3. Pilot Studies and Simulations for Sample Size Calculations

When no reasonable assumptions can be made about the means and standard deviations of indicators of interest, a small-scale pilot study may help in estimating them. Pilot studies confer numerous other advantages by giving the evaluator a preview of how the research will be carried out in practice. This can allow her to see logistical and other practical issues that arise in collecting data and perhaps refine the research questions based on evidence from the pilot study. This should include an estimate of the non-response rate and whether it may be linked to selection bias.

To get a sense of how the inclusion of covariates in a multivariate regression may influence sample size and statistical power, the researcher can use Monte Carlo simulations. Using existing secondary data—perhaps from a survey on a similar population in a similar setting (like a neighboring country) or data from a pilot study—that approximate the data to be collected in the field, simulations have the advantage of allowing the evaluator to get a feel for how changes in various parameters (effect size and standard deviation), and in control variables included the regression model, can affect statistical power and sample size. (For more on using simulations to determine sample size, see Lenth [2001].)

D. Constructing Valid Comparison Groups

The manner of selecting the comparison group for the sample can greatly influence the ease and rigor of the subsequent analysis. How to select a comparison sample depends largely on program placement and how beneficiaries were selected. Randomized allocation of the program among eligible beneficiaries is the simplest case: the researcher simply includes eligible units (individuals, villages, etc.) that were not (yet) selected to participate. Targeted programs without randomized assignment and programs that rely on some form of self-selection of participants are more complicated, and we discuss some ideas for selecting comparison groups in these contexts here.

If existing data were collected prior to project implementation and the evaluation will use propensity score matching analysis, the secondary data can be used to estimate the probability of participating in the program. Having completed this first stage of the propensity score matching, comparison observations can be selected to best match treatment observations. Likewise, if the evaluation plans to use a regression discontinuity design, the sample can be selected to capture treatment and comparison observations closest to the cut-off criteria. Such matching must be done at the level of observation at which program participation was assigned. For programs that selected communities to receive an intervention, matching must be done using community-level variables. The same applies for programs that selected beneficiaries at the facility, household, or individual levels. If secondary data are not available, or if selection criteria are not clearly defined and strictly followed, these methods will not be applicable. Nevertheless, for any targeted intervention, it is worthwhile to try to structure the sample to capture the targeted groups. This will avoid dropping observations at the time of analysis, for lack of common support.[41] Focusing a sample in these ways will not increase statistical power, but it can help avoid losing power later due to discarded observations.

For programs that have an element of demand-driven selection, constructing a valid comparison group can be even trickier. This is because some unobserved factors influence some people to choose to participate while others choose not to. This introduces a source of bias which is difficult to control for in the analysis, since it stems from

[41] Common support refers to the range of estimated propensity scores for which treatment and comparison observations are both found. Observations outside the region of common support are those for which there are no comparable observations in the other group, and so must be dropped to avoid biased results.

unobserved factors.[42] If, however, a program was not administered to all eligible applicants, and if admission or exclusion was not systematically determined (i.e., effectively random), the people who applied to the program but did not participate make a natural comparison group. It is essential that participants were not chosen from among eligible applicants based on any criteria. For instance, if 1000 eligible people applied for a labor training program that could only accommodate 500, and 500 names were essentially drawn from a hat, the 500 left out make a valid control group. If, however, the 500 beneficiaries were selected because they were the poorest half, or because they had more children than the others, or for any other reason, the applicants who were left out do not make a comparable comparison group. Such rules would perhaps permit a regression discontinuity design. But where the selection was more ad hoc, constructing a valid control group is more difficult.

E. Panel Data [43]

Panel data are two or more repeated observations on the same units—be they individuals, households, villages, etc.—at different points in time. Constructing (or acquiring) panel data requires special considerations. (In the case of collecting new primary data to construct a panel dataset, it requires implementing the recommendations in this section. In the case of acquiring existing panel data, it requires verifying that they were done.) The first requirement is to re-sample the same observations from the old sample, to the extent possible. A decision must be made as to which unit of observation should be the same over time. For most household surveys, this is the household itself, or one or more key individuals within it. So if they move, the second round survey should aim to find them in their new dwelling. Yet for some research topics, the dwelling and perhaps lands attached to it are a primary focus, and so the survey may resample the same places, regardless of whether the inhabitants are the same as before or not.

The second requirement is to repeat the questions from the original versions of the questionnaire verbatim in the latter versions of the questionnaire, so as to permit valid comparisons over time. One can add questions to later versions of questionnaires, but changing the wording, format, or even sometimes the order of questions may invalidate their comparison over time.

If an evaluation team is planning on constructing a new panel dataset from scratch, several considerations come into play in the implementation of the first round survey. These will greatly facilitate the later round(s) of data collection and combining the data into the full panel. One is to assign and record unique identification codes to each household and to each member within it. Household codes should be retained in subsequent survey rounds to make clear the identification of households over time. Person ID codes should be mapped from one round of data collection to the next, by recording the person ID codes used in the previous round of data collection in the household roster along with the new codes for the current survey round. To see examples

[42] Panel data (see section E on this page) allows methods to be used which control for time-invariant factors.

[43] Glewwe and Jacoby (2000) discuss issues related to constructing and collecting panel data in much more detail.

of how matching observations across rounds of panel data collection has been managed in practice, readers can find Basic Information Documents that describe procedures used to collect household panel data in Côte d'Ivoire and in Nepal at http://www.worldbank.org/lsms/guide/select.html.

Another consideration is to record GPS coordinates for each household, to reduce ambiguity and difficulty in locating the same households later. If this is not practicable, the exact location of each dwelling should be recorded in as much detail as possible, including mapping the location of sampled households, to make it easy to find them again in the future. Likewise, recording information that will help track down people who move can be important in reducing attrition from one round to the next. For example, surveys can include questions to record the contact information of neighbors or others who will be most likely to know where a family has moved to if they do relocate. For programs that might induce migration, it may be even more important to take these kinds of steps to enable the research team to track down people who move around between survey rounds.

It is worth mentioning here that panel data samples become progressively less representative over time. New households created after the time when the initial panel sample was selected are excluded by construction. For this reason, most panel datasets that run over a few years are constructed as rotating panels, where only a portion of the old observations are revisited, and new observations are added in. This helps maintain a representative sample while still allowing for tracking some households over longer periods of time.

Finally, for the purposes of a program evaluation, it is necessary to track how people move into and out of the program over time. For interventions that are not available to all people in a certain area, it is necessary to include a participation indicator in the data. In latter rounds of data collection, it is additionally necessary to record how long a beneficiary has been participating in the program, so that the date of entry into the program can be established with reasonable precision ("reasonable" being context-dependent).

F. Questionnaire Length

Because collecting primary data involves sending survey enumerators (interviewers) out into the field, the time it takes to administer each questionnaire influences the sample size that can be generated within a fixed budget. If statistical power calculations show that the necessary sample size is larger than what seems feasible with available resources, reducing the length of the questionnaire can increase the sample size that can be generated with a given budget, because each survey enumerator can interview more people or households each day. When conducting primary data collection on a relatively tight budget, there may be a trade-off between the scope of research and the statistical precision. It may be worth narrowing the scope of research by collecting fewer indicators on more observations to allow analysis that still gives reasonably precise results.

As was discussed in subsection VI.C. above, cluster effects need to be taken into account in sample size calculations. Increasing the number of interviews each enumerator can complete in a day of fieldwork may not improve statistical power much if households are quite similar within communities (intra-cluster correlation is high). If this is the case, it might be better to use the time savings from a shorter questionnaire to travel to more PSUs, rather than interviewing more households in each PSU.

G. Documentation of Sample Selection and Field Implementation

Producing detailed written records of how sample selection was determined and how it was modified during fieldwork (e.g., replacement of households that could not be located or refused to be interviewed) is a key component of quality data management. This is indispensable for researchers to be able to use the data properly and conduct rigorous analysis. Without proper documentation describing all facets of how the data was generated, the quality of research based on that data is diminished.

VII. A Final Note: Making Data Available

Whenever possible, data should be made publicly available with evaluation results and reports. One group's primary data may serve as another group's secondary data, reducing costs of future evaluations. Moreover, the data can be used by other researchers to review the methods and confirm the results of an evaluation. Publicly available data can even expand the scope of an evaluation by allowing researchers to use the data to investigate further topics not addressed by the original evaluation team. (This is exemplified by the ever-expanding number of studies which have used the data collected to evaluate Progresa/Oportunidades in Mexico.) Before making data available however, it is important to make sure that issues surrounding data confidentiality have been satisfactorily resolved.

Suggested Further Resources

Baker (2000) provides a very accessible introduction to the interrelation of data and methods in program evaluation. Her book also provides a number of case studies that describe how data were employed in actual evaluations. These examples are useful in getting a sense of how eclectic evaluators often need to be regarding the data they use, and how data availability impacts the analytical methods evaluators can use in their studies. Her book is available online at http://go.worldbank.org/8E2ZTGBOI0.

Grosh and Muñoz (1996) offer a comprehensive manual for designing and implementing large-scale household surveys, focusing in particular on how Living Standards Measurement Study surveys are implemented. Their paper is a good resource for users of LSMS data and a font of practical advice for those planning household surveys. It is available online at http://go.worldbank.org/T81ZN6GZE0.

Iarossi (2006) provides a valuable resource on generating and using survey data, including discussion of questionnaire design, sampling, survey response psychology, and overall management of the data generation process.

United Nations (2005) is an extensive text on conducting household surveys in developing and transition countries, including sample design, survey implementation, non-sampling errors, survey costs, and analysis of survey data. It is available online at http://unstats.un.org/unsd/hhsurveys/

Useful websites related to data for impact evaluations:

http://www.internationalsurveynetwork.org/home maintains a list of completed and planned surveys for many countries, as well as access information for a survey data management software package.

http://www.worldbank.org/impactevaluation has many resources on various aspects of doing impact evaluation, including management issues and analytical methods.

http://www.worldbank.org/lsms has many examples of survey questionnaires and manuals used in multi-topic household surveys in many countries.

http://checkyourdata.com/ has an online utility where data can be checked for inconsistencies using Benford's Law.

Bibliography

Annan, Jeannie, Christopher Blattman, and Roger Horton (2006). "The State of Youth and Youth Protection in Northern Uganda: Findings from the Survey for War Affected Youth," September 2006 Phase 1 Final Report for UNICEF Uganda, available online at http://www.sway-uganda.org/.

Baker, Judy (2000). *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: World Bank, available online at http://go.worldbank.org/8E2ZTGBOI0.

Bamberger, Michael, Jim Rugh, Mary Church, and Lucia Fort (2004). "Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints," *American Journal of Evaluation* 25(1): 5–37.

Bamberger, Michael (2006). *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints*. Washington, DC: World Bank.

Castelloe, John and Ralph O'Brien (2001). "Power and Sample Size Determination for Linear Models," Paper 240-26, Proceedings of the Twenty-Sixth Annual SAS Users Group, available online at http://www2.sas.com/proceedings/sugi26/p240-26.pdf.

Cochran, William (1977). *Sampling Techniques* (3rd ed.). Santa Barbara: John Wiley & Sons, Inc.

Coudouel, Aline, Jesko Hentschel, and Quentin Wodon (ca 2001). "Chapter 1: Poverty Measurement and Analysis" in *PRSP Sourcebook Volume 1: Core Techniques and Cross-Cutting Issues*, Washington, DC: World Bank, available online at http://povlibrary.worldbank.org/files/5467_chap1.pdf.

Deaton, Angus (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: Johns Hopkins University Press.

de Negri, João Alberto, Mauro Borges Lemos, and Fernanda de Negri (2006a). "Impact of R&D Incentive Program on the Performance and Technological Efforts of Brazilian Industrial Firms," Inter-American Development Bank, Office of Evaluation and Oversight, Working Paper OVE/WP-14/06, December 2006, available online at http://www.iadb.org/ove/Documents/uploads/cache/907639.pdf.

de Negri, João Alberto, Mauro Borges Lemos, and Fernanda de Negri (2006b). "The Impact of University Enterprise Incentive Program on the Performance and Technological Efforts of Brazilian Industrial Firms," Inter-American Development Bank, Office of Evaluation and Oversight, Working Paper OVE/WP-13/06, December 2006, available online at http://www.iadb.org/ove/Documents/uploads/cache/907638.pdf.

Duflo, Esther (2001). "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review* 91(4): 795-813.

Dupont, William and Walton Plummer (1990). "Power and Sample Size Calculations," *Controlled Clinical Trials* 11: 116-128.

Eng, John (2003). "Sample Size Estimation: How Many Individuals Should Be Studied?" *Radiology* 227: 309-313, available online at http://radiology.rsnajnls.org/cgi/content/abstract/227/2/309.

Eng, John (2004). "Sample Size Estimation: A Glimpse beyond Simple Formulas," *Radiology* 230: 606-612, available online at http://radiology.rsnajnls.org/cgi/content/abstract/230/3/606.

Ferraz, Claudio and Frederico Finan (2007). "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes," May 2007, available online at http://www.econ.ucla.edu/ffinan/Finan_Audit.pdf.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky (2005). "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy* 113(1): 83-120.

Glewwe, Paul and Hanan Jacoby (2000). "Chapter 23: Recommendations for Collecting Panel Data," in Grosh, Margaret and Paul Glewwe, eds., *Designing Household Survey Questionnaires for Developing Countries*. Washington, DC: World Bank.

Grosh, Margaret and Juan Muñoz (1996). "A Manual for Planning and Implementing the Living Standards Measurement Study Survey," World Bank LSMS working paper 126, available online at http://go.worldbank.org/T81ZN6GZE0.

Heckman, James, Robert Lalonde, and Jeffrey Smith (1999). "Chapter 31: The economics and econometrics of active labor market programs," in Ashenfelter, Orley C. and David Card, eds., *Handbook of Labor Economics* 3(1): 1277-2097.

Iarossi, Giuseppe (2006). *The Power of Survey Design*. Washington, DC: World Bank.

Judge, George and Laura Schechter (2006). "Detecting Problems in Survey Data using Benford's Law," working paper, available online at http://www.aae.wisc.edu/schechter/benford.pdf.

Kraemer, Helena Chmura and Sue Thiemann (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park: Sage Publications.

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

Leamer, Edward (1983). "Let's Take the Con Out of Econometrics," *American Economic Review* 73(1): 31-43.

Lenth, Russell (2005). "Russ Lenth's power and sample-size page," available online at http://www.cs.uiowa.edu/~rlenth/Power/.

Lenth, Russell (2001). "Some Practical Guidelines for Effective Sample Size Determination," *The American Statistician* 55(3): 187-193.

Lohr, Sharon (1999). *Sampling: Design and Analysis*. Pacific Grove, CA : Duxbury Press.

Lokshin, Michael and Ruslan Yemtsov (2003). "Evaluating the impact of infrastructure rehabilitation projects on household welfare in rural Georgia," World Bank Policy Research Working Paper 3155, available online at http://go.worldbank.org/DK6VLBJL60.

Lokshin, Michael and Ruslan Yemtsov (2005). "Has Rural Infrastructure Rehabilitation in Georgia Helped the Poor?" *World Bank Economic Review* 19(2): 311-333, available online at http://wber.oxfordjournals.org/cgi/reprint/19/2/311.

Mahalanobis, Prasanta Chandra (1946). "Recent experiments in statistical sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society* 109(4): 326-370.

Prennushi, Giovanna, Gloria Rubio, and Kalanidhi Subbarao (2002). "Chapter 3: Monitoring and Evaluation" in *PRSP Sourcebook Volume 1: Core Techniques and Cross-Cutting Issues*, Washington, DC: World Bank, available online at http://povlibrary.worldbank.org/files/4480_chap3.pdf.

Rao, Vijayendra and Ana María Ibáñez (2003). "The Social Impact of Social Funds in Jamaica: A Mixed-Methods Analysis of Participation, Targeting, and Collective Action in Community-Driven Development," World Bank Policy Research Working Paper 2970, available online at http://go.worldbank.org/81LTI3JRA0.

Ravallion, Martin (2001). "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation," *World Bank Economic Review* 15(1): 115-140, available online at http://wber.oxfordjournals.org/cgi/content/abstract/15/1/115.

Ravallion, Martin (2006). "Evaluating Anti-Poverty Programs," forthcoming in *Handbook of Development Economics Volume 4*, edited by Robert E. Evenson and T. Paul Schultz, Amsterdam: North-Holland, available online at http://siteresources.worldbank.org/INTISPMA/Resources/383704-1130267506458/Evaluating_Antipoverty_Programs.pdf.

Savedoff, William, Ruth Levine, and Nancy Birdsall (2006). "When Will We Ever Learn? Improving Lives Through Impact Evaluation," Washington, DC: Center for Global Development, May 2006 final version available online at http://www.cgdev.org/content/publications/detail/7973. (September 15, 2005 Consultation Draft available online at http://www.cgdev.org/doc/eval%20gap/draft9.15.05web.pdf.)

United Nations (2005). *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations, available online at http://unstats.un.org/unsd/hhsurveys/.