IDB
Development Effectiveness

Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-123
May 2010

Improving the Quality of Data and Impact-Evaluation Studies in Developing Countries

Guy Stecklov
Alex Weinreb

Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank

Improving the Quality of Data and Impact-Evaluation Studies in Developing Countries

Impact-Evaluation Guidelines

Guy Stecklov
Alex Weinreb

Inter-American Development Bank
2010

Guy Stecklov. Hebrew University, Jerusalem. guy.stecklov@mail.huji.ac.il

Alexander Weinreb. University of Texas, Austin. aweinreb@prc.utexas.edu

Improving the Quality of Data and Impact-Evaluation Studies in Developing Countries

Abstract

Guy Stecklov* and Alex Weinreb**

While the science of program evaluation has come a tremendous distance in the past couple of decades, measurement error remains a serious concern and its implications are often poorly understood by both data collectors and data analysts. The primary aim here is to offer a type of "back-to-basics" approach to minimizing error in developing country settings, particularly in relation to impact-evaluation studies. Overall, the report calls for a two-stage approach to dealing with mismeasurement. In the first stage, researchers should attempt to minimize mismeasurement during data collection, but also incorporate elements into the study that allow them to estimate its overall dimensions and effects on analysis with more confidence. Econometric fixes for mismeasurement—whose purview is limited to a smaller subset of errors—then serve as a secondary line of defense. Such a complementary strategy can help to ensure that decisions are made based on the most accurate empirical evaluations.

The main body of the report includes four main sections. Section two discusses in detail many of the problems that can arise in the process of data collection and what is known about how they may affect measurement error. Section three provides a basic introduction to statistical—particularly econometric—methods that have been developed and used to help avoid the most problematic effects of mismeasurement. Section four offers an alternative approach to dealing with measurement error—one that focuses on reducing error at source. It offers pointers to current "best practice" on how to reduce measurement error during data collection, especially as to how those methods relate to evaluation research, and how to incorporate elements into research design that allow researchers to estimate the dimensions of error. Section five focuses on the role of incentives as one possible approach to shifting one particular aspect of error. It uses data from the PROGRESA program to evaluate indirectly the impact of incentives on certain aspects of data quality in Mexico. The report concludes with a short summary and includes a list of ten steps that can be taken to reduce measurement error at source.

JEL Classification: C8, C9, I3
Keywords: development effectiveness, impact evaluation, randomization, survey design, measurement error.

* Associate Professor, Department of Sociology and Anthropology Hebrew University, Jerusalem. E-mail: guy.stecklov@mail.huji.ac.il

** Associate Professor, Department of Sociology Faculty Associate, Population Research Center University of Texas, Austin. E-mail: aweinreb@prc.utexas.edu

Table of Contents

Foreword

As part of its efforts to improve the development effectiveness of IDB-funded projects, the Strategy Development Division supports the IDB's operational divisions and its country partners in designing and implementing rigorous impact evaluations of those projects.

There are numerous challenges in conducting impact evaluations. Some of these challenges, such as issues in data collection and the methods of evaluation, are more general issues and relevant for most impact evaluations, while other challenges are specific to the type of project being evaluated.

To help address these challenges, the Strategy Development Division has put together this series of guidelines on doing impact evaluations. Some of the papers in the series provide general guidance on how to do impact evaluations, while others focus on providing insights into conducting impact evaluations of certain types of projects.

These guidelines aim to facilitate rigorous evaluation of the impact of projects funded by the IDB. Only through such an approach can lessons be learned from IDB-funded projects and improvements in development effectiveness be made.

Carola Alvarez
Chief
Strategy Development Division

Section 1. Introduction

The science of program evaluation has come a tremendous distance in the past couple of decades. Earlier strategies which relied on limited and informal assessments have been gradually supplanted by the integration of ongoing impact evaluation into original project designs. As the value of this integrated approach has grown clearer, it has become increasingly desirable that program evaluations be integrated into programs at the design stage. At the same time, a widely acknowledged gold standard for evaluation has also emerged: the experimental approach, wherein people, communities, or organizations are randomly assigned to treatment and control groups, enabling researchers to easily identify the causal impact of the program. Yet, despite these advances, and the obvious strength of the experimental approach in identifying causal effects, hurdles continue to impede efforts to understand program impact. This report focuses its attention on one of them, the quality of data.

The quality of data is an essential element in all impact-evaluation efforts. This includes both the randomized, gold-standard evaluation as well as in the more traditional nonrandomized evaluation. Mismeasurement in fact can be a major Achilles' heel in any statistical approach that utilizes empirical evidence, particularly in developing countries. This is true even in highly regarded data-collection processes, such as the World Bank's Living Standards Measurement Surveys (LSMS) or the Demographic and Health Surveys (DHS). Although both of these make excellent documentation available about certain elements of data collection, other elements receive far less attention. In part this stems from a simple problem: there is only a limited methodological literature on which scientifically-grounded decisions regarding data collection can be based. And with very few exceptions, that literature is based on methodological experiments conducted in a small number of developed societies. The extent to which these validation studies can inform methodological practice in developing countries, which are structurally dissimilar, remains unclear.

Understanding the range of dimensions across which these structural differences exist is important. Take, for example, a group of prospective respondents enrolled in a comparative evaluation study in a developed and a developing country. If chosen randomly from the general population, they will vary not only in their age, wealth and education, but also in the intensity of their family-based interaction, their level of linguistic and ethnic heterogeneity, the strength of their local patronage systems, their cultural understanding of "confidentiality" (which underlies

accurate reporting) and so on. Each of these differences—and others like them—has implications for how researchers should go about collecting data. More importantly, cumulatively, they force researchers to question the quality of the empirical evidence used to substantiate claims about a given program's success or failure. Such questions are all the more vital given the rapid and ongoing expansion of efforts to evaluate important development interventions.

This report aims to help build stronger capacity within developing countries to collect high-quality data for evaluation purposes and to recognize where problems may occur.

This involves an examination of which data-collection strategies have been validated and may be best utilized in developing countries. Currently available work on data-collection methods for developing country surveys—based solely on recommendations by experienced field researchers, not experimental design—are out-of-date (e.g., Caldwell et al., 1970; Kearl, 1976; Casley and Lury, 1981; Bulmer and Warwick, 1983/1993). They have missed important developments in data-collection methodology in developed countries, as well as the more formal, methodological literature on data collection in developing country settings that has emerged over the last several years. The same is true, albeit to a more limited degree, in relation to information on data collection in developing countries, although a large number of studies has been published by the World Bank.

The primary aim here, therefore, is to offer a type of "back-to-basics" approach to minimizing error in developing country settings, particularly in relation to impact-evaluation studies. The report describes a range of specific steps that can be taken during the process of data collection that can either help researchers avert or reduce mismeasurement, or allow them to estimate its overall dimensions and effects on analysis with more confidence. These steps—it must be emphasized—are not intended to be substitutes for the established repertoire of statistical and econometric methods that are already at the disposal of evaluation researchers. The latter have considerable advantages. They are widely taught, described at length in all introductory econometrics texts, and tend to have estimation procedures bundled with standard statistical software. They can also be applied post hoc, that is, after the data have been collected. On the other hand, these methods also have some notable disadvantages. Chief amongst these, they require strong assumptions which may only be partly appreciated by their users. As a result, use of these methods may insert more distortion into the findings than is found in the unaltered data (Bound, Brown and Mathiowetz, 2001).

For these reasons, this report favors a balanced approach to minimizing error. Whether one is dealing with research in general or evaluation research in particular, it can be referred to as a complementary or tandem strategy. It begins with an attempt to minimize mismeasurement during the data-collection stage. And it continues into the analytic stage, introducing econometric fixes as a type of secondary line of defense whose purview is limited to a smaller subset of errors.

The report is divided into four main sections.

Section 2 discusses in detail many of the problems that can arise in the process of data collection and what is known about how they may affect measurement error. It also provides a list of selected references for further reading.

Section 3 provides a basic introduction to the statistical methods that have been developed and used to help avoid the most problematic effects of mismeasurement.

Section 4 offers an alternative approach to dealing with measurement error—one that focuses on reducing error at source. It also offers pointers to current "best practice" on how to reduce measurement error, especially as to how those methods relate to evaluation research.

Section 5 focuses on the role of incentives as one possible approach to shifting one particular aspect of error. It explores theories that may make it easier to predict the impact of incentives on measurement error and uses data from the PROGRESA program to evaluate indirectly the impact of incentives on certain aspects of data quality in Mexico.

These four sections are then followed by conclusions including ten steps that can be taken to reduce measurement error at source.

Readers most likely to profit from this report are those directly involved in evaluation research in developing countries, whether in an impact evaluation that is based on a randomized treatment, or in a more traditional approach where randomization is not possible. Evaluation is such a critical enterprise, even when using less than ideal data, researchers should not be put off by the challenge. In addition, it is hoped that developing country researchers in general will also benefit. However, since the literature on both data collection and econometric approaches to error is enormous and constantly being updated, this report is not an exhaustive discussion of these issues. Nor is it the final word. But it does cover a range of vital issues. Moreover, it is hoped that the overall summary, synthesis, and recommendations will make the report useful. For not only does it update past syntheses of data-collection methods in developing country

surveys; it also offers readers some basic ground rules about what to do, when, and how, in each case pointing them to secondary literature for more detailed review.

The tendency to ignore data-collection issues in large-scale surveys and evaluation research in developing countries is a relatively new one. Studies of, or at least reflections on, data-collection problems were considerably more common during the early days (1950s and 1960s) than during the last few decades.

Aside from the book-length collections on field methods noted above, some informative papers from those early days include:

Back, K.W., and J.M. Stycos. 1959. The Survey under Unusual Conditions: Methodological Facets of the Jamaica Human Fertility Investigation. Cornell University, Ithaca, New York, United States: Society for Applied Anthropology.

Choldin, H,M., A.M. Kahn and B.H. Ara. 1967. "Cultural Complications in Fertility Interviewing". *Demography* 4(1):244-52.

Mauldin, W.P. 1965. "Application of Survey Techniques to Fertility Studies". In: M.C. Sheps and J.C. Ridley, editors. *Public Health and Population Change: Current Research Issues*. Pittsburgh, Pennsylvania, United States: University of Pittsburgh Press.

Mitchell, R.E. 1965. "Survey Materials Collected in the Developing Countries: Sampling, Measurement and Interviewing Obstacles in Intro- International Comparisons". *International Social Science Journal* 17:665-85.

Poti, S.J., B. Chakraborti and C.R. Malaker. 1962. "Reliability of Data Relating to Contraceptive Practices". In: C.V. Kiser, editor. *Research in Family Planning*. Princeton, New Jersey, United States: Princeton University Press.

Rudolph, L. and S.H. Rudolph. 1958. "Surveys in India: Field Experience in Madras State". *Public Opinion Quarterly* 33:235-44.

Stycos, J.M. 1960. "Sample Surveys for Social Science in Underdeveloped Areas". In: R.N. Adams and J. Preiss, editors. *Human Organization Research*. Homewood, Illinois, United States: The Dorsey Press.

Section 2. The Problems of Measurement Error

A. The centrality of measurement in the evaluation process

Impact evaluation has become a cornerstone of public and nonpublic development programs. The availability of data—good data—lies at the core of these efforts. Some data may be obtained through routine data-collection systems, particularly where there is ongoing monitoring. Evaluation studies in general, however, greatly benefit from random-sample surveys. This is particularly the case where the evaluation efforts are focused on assessing the impact of programs on the population. The reason is that surveys provide researchers and policymakers with data across the widest range of people, behaviors and activities represented in the target population. With regard to behaviors and activities, this includes data used to measure specific program objectives, such as reducing poverty, raising child school attendance rates, or raising female contraceptive use. Crucially, however, it also includes data on other behaviors or related variables (e.g., attitudinal change), which may not be the intended program target but which may reflect unintended externalities produced by development programs. These unintended outcomes are also important for evaluation (whether or not they are consistent or inconsistent with development goals). High-quality survey-data collection, in short, can dramatically increase the efficacy of program impact evaluation in developing countries.

i. What exactly is measured in surveys?

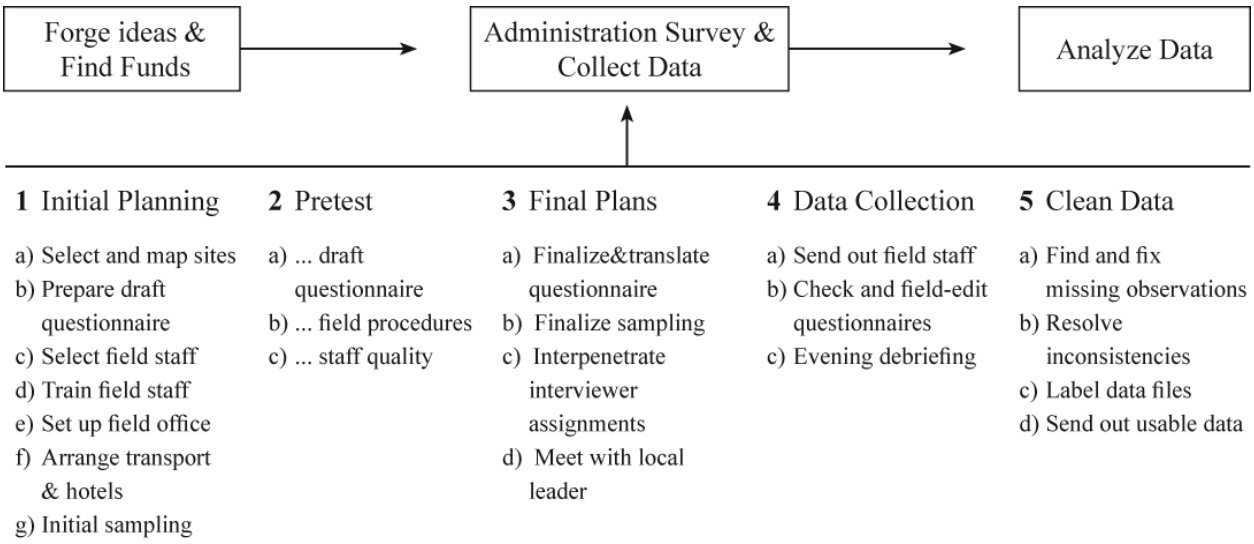
Sample surveys enable researchers to collect information about a population without having to collect data from every member of that population. The specific "units of analysis" in that population depend on the research question. They can be individuals, families, communities, firms, or any other unit. As should be clear from the phrase "sample surveys," surveys are intrinsically related to samples and sampling, some basic principles of which are described below.

As described in Figure 2.1, there are essentially five stages involved in the administration of a survey. All five fall between the initial "forging ideas and finding funds" stage and the final "analyzing data" stage. These stages, particularly the last, tend to be covered thoroughly during graduate-school training. Survey administration, in contrast, does not.

Broadly speaking, the five stages in the administration of a survey are: (i) initial planning, (ii) pretesting, (iii) finalizing data-collection plans, (iv) actual data collection, and (v) data cleaning. As will become clear in the remainder of this section and in the following sections, errors in surveys can enter the data during all five stages. Survey error, in other words, has its roots in bad initial planning (both administrative and scientific), inadequate pretests (of personnel and field procedures as well as questions), rushed final planning, and so on.

The key point here is that the collection of high-quality data is more complicated than it may appear, not only administratively, but also scientifically. The array of problems that will be described in the following pages—scientific, administrative, interpersonal, political—must either be avoided or, since this is not always possible, the research design must be manipulated in order to allow researchers to evaluate their impact on the data. In the absence of such steps, it is impossible for researchers to identify the sources of variance in their data.

Figure 2.1 A Five-stage Roadmap of Survey Administration



The collection of high-quality survey data begins with the careful definition of core constructs and specification of units of analyses. The core constructs of interest—that is, the main concepts that a researcher seeks to measure and understand—must be carefully specified in order to make certain that questions are asked that cover the domain of interest. No less important, the units of analysis need to be decided before any survey is put to the l, so that the

sampling strategy and data-collection approach are appropriate. For example, even seemingly straightforward variables such as "education," often used to capture human capital, can be measured in a number of ways (e.g., total years of schooling, highest level attained, quality of education, actual skills acquired). Each has somewhat different implications for interpretation. Researchers must therefore first clearly define what it is they want to measure. Such careful specification is necessary before the level of measurement error can begin to be gauged.

The unit of analysis also needs to be carefully specified, and appropriate to the core constructs. For example, a single individual might be able to provide information about themselves and, somewhat less reliably, about all members of their household. Their ability to provide information about a larger collective like an organization may be even more limited. But such information may be necessary given the evaluation goals and data-collection opportunities. Data-collection requirements, in other words, vary considerably where programs are focused on individuals, households, or organizations. And the expected types and levels of mismeasurement will vary accordingly.

Either way, these primary stages of measurement-related decisions are not the focus here. Rather, as implied in Figure 2.1, the focus here on measurement and mismeasurement refers primarily to what happens at the next stage, that is, after decisions about what data to collect have already been made. Specifically, how does one go about collecting the desired (or required) data with minimal error?

ii. Types of mismeasurement in surveys

The critical point is this: measurement error can affect data—and consequently affect the conclusions reached from survey-based evaluations—at every stage of the data-collection process, that is, during any of the five stages in the lower panel of Figure 2.1.

Over the last several decades, those stages have been delineated in a number of ways. Perhaps the most common approach is to distinguish between "sampling error" and "nonsampling error" (e.g., Feinberg, 1990). Another approach, suggested by Groves (1989), distinguishes "errors of omission" (largely sampling errors) with "errors of commission." In the present discussion, a slightly different approach is used. Following a number of scholars, the approach presented here is based on the Total Survey Error Model (TSEM), the origin of which

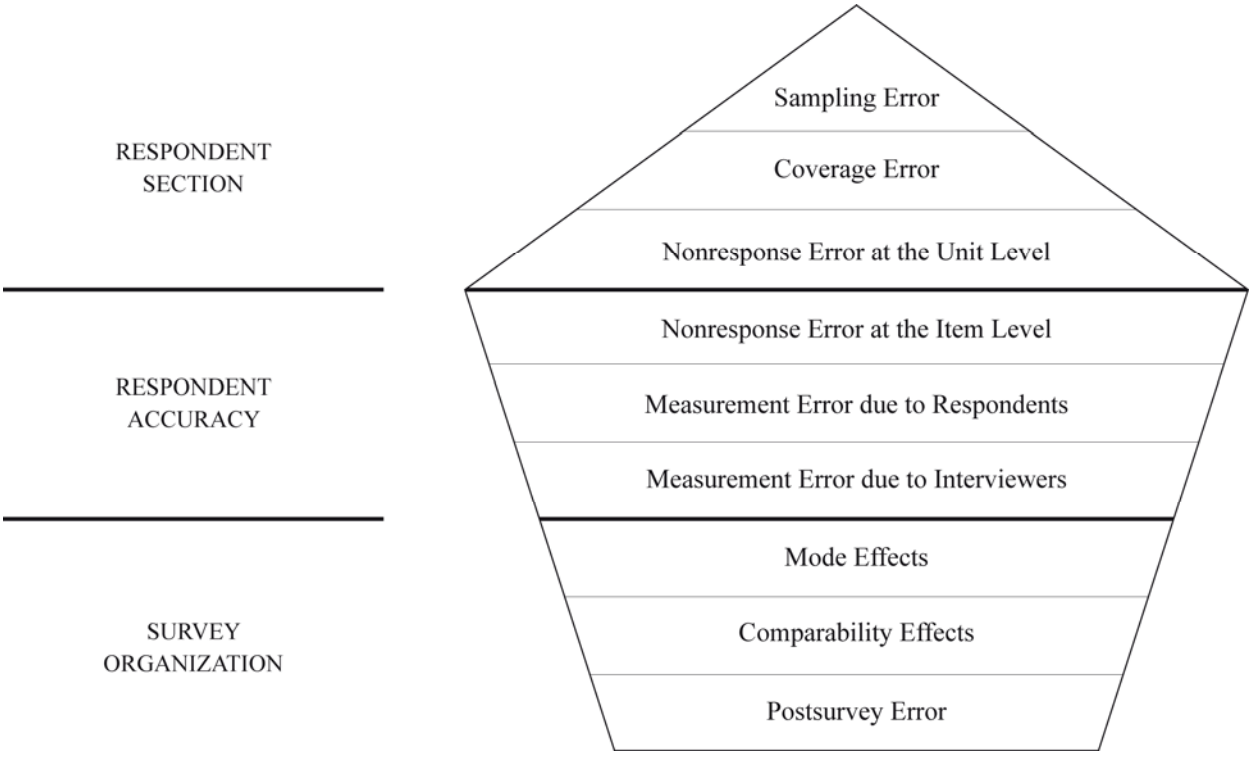
is typically linked to Deming's (1944) foundational paper, and whose primary aim is to elucidate the multiple sources and dimensions of error.

Figure 2.2 provides a graphical representation of the TSEM. It contains nine distinct levels or stages at which measurement error can be introduced. These nine levels can be lumped into three general categories of mismeasurement issues, related to respondent selection, accuracy of respondents' answers, and survey administration.

The model depicted in Figure 2.2 has informally been described as an "iceberg" model. The reason is that only the top levels are actually visible above the surface and commonly acknowledged by most researchers. Within the top level, "sampling error" is by far the most widely recognized of the errors.

Each of these sources of error—those lurking below as well as those that are fully visible—is now described, in order to show how they can threaten survey data quality. Section 4 then describes steps that researchers can take to avoid or minimize their effects. The sources of error are covered in the top-to-bottom order of Figure 2.2, beginning with sampling error.

Figure 2.2 Total Survey Error Model, as interpreted in Weisberg (2005)



Ideally, all variables that a researcher decides to measure—when carefully defining core constructs and specifying units of analyses—would be measured perfectly. In the real world, combinations of sampling vs. nonsampling errors or errors of omission vs. commission make this impossible.

Sampling error: Sampling error refers to the difference between a statistic measured from a sample of the population and the same statistic derived from data on the entire population. It is convenient to think of each sample as one among a huge number of possible samples that could be drawn from the population, with each being slightly different. For example, mean years of schooling may be 7 in the total population, 6 in one sample survey, and 4.5 in a second survey. Sampling error in this case is clearly greater in the second survey.[1]

With regard to sampling error, the central problem for researchers is that the extent of sampling error is typically unknown. Instead of estimating it directly, researchers rely on probability theory to estimate its dimensions. This, in fact, is one of the primary functions of sample-design tools such as "power calculations." Such calculations demand that users specify an assumed distribution for sample statistic x and a desired level of confidence in the relationship between x and population statistic X. The power calculation then tells researchers how many people are needed in the sample. Of course, such calculations become more complicated as one moves away from a simple random sample—see footnote—but the basic principle remains: by knowing the level of sampling error, researchers can generate confidence intervals around sample estimates, allowing them to generalize from those sample estimates to the true population statistic.

[1] Sampling provides a way to characterize a population without having to collect data from every member of that population, whether those members–the "units of analysis" referred to above–are individuals, families, communities, firms, or some other unit. Underlying this aim is the principle of "generalization." One can directly generalize from a sample to the total population if and only if sampled observations are purposively random. While nonrandom sampling mechanisms may also generate a truly random sample, the point is that they do not necessarily do so. This makes it difficult to know how representative those samples are.

Survey research, therefore, systematically randomizes. Over the last century or so, a number of approaches have developed. The simplest is the classic one-stage simple random sample (SRS). A slightly more complicated approach is to use sample weights in order to generate a sufficient subsample from smaller populations of interest (e.g., 5 percent of the total population which researchers want to constitute 15 percent of the sample). This approach generates what is known as a stratified sample. Yet another approach, increasingly popular, is a cluster-sampling strategy. In this case, a sample of clusters is selected (from a universe of clusters), and all eligible individuals within those clusters are interviewed. Finally, the most popular contemporary approach in large, complex surveys is the multi-stage sampling strategy. Here, clusters are selected as per the cluster-sampling approach, but then only a randomly selected subsample of eligible units within those clusters is selected.

Even if at least one of these inputs is guesswork—researchers sometimes have little-to-no idea what the distribution of X is—sampling errors are the most familiar to researchers. They are covered in all introductory statistics courses. Moreover, they are easy to fix: no matter what the sampling strategy, researchers can reduce sampling error by increasing sample size.

If sampling error is familiar to researchers and easy to fix, nonsampling error is the opposite on both counts. This is deeply problematic since data-collection methodologists have shown that nonsampling errors—of the type that are listed in the lower levels of Figure 1—are often larger than sampling errors, in some cases much larger (e.g., Powell and Pritzker, 1965). This is even the case on variables as simple to measure as educational attainment, and even in developed country settings where norms of survey-based data collection are more established (Bailar, 1976). This has significant implications for the collection of high-quality data.

Coverage error, the next layer in the respondent-selection category, refers to the error introduced when the sampling frame for the survey does not represent the intended population of interest. In this case, however good the sample is technically, the actual respondent population will not provide a good basis for generalization to the population as a whole. This problem can be quite severe where sampling frames—often national censuses—are unreliable or difficult to access. For example, developing country censuses are often inaccurate at the local (e.g., district) level, since data-smoothing techniques tend to be administered on national-level data only. In addition, since such data are highly politicized—particularly in heterogeneous polities—census data in a number of developing countries are either completely out of date (e.g., Lebanon's last was in 1932; Iraq's was in the 1987), or not gathered for long periods (e.g., a 25-year lapse prior to Sudan's 2008 census, 18 years in Cameroon prior to 2005 census). Even if censuses are held, results are often deemed too politically sensitive for public release (the case in many countries divided along ethnic and religious lines). In each of these cases it makes it difficult, or even impossible, to draw a reliable sampling frame from the data, and any such attempt is liable to generate considerable coverage error. This is particularly true in settings in which there are significant differences in population growth rates at the subnational, regional level, as is typical of many developing countries in which high-impact evaluation research is, or needs to be, conducted.

Even if the sampling frame is accurate and up-to-date, *considerable nonresponse error at the unit level* can occur where surveys do not successfully obtain interview responses from a high proportion of selected individuals or households—depending on the unit of analysis—or where there is evidence of selectivity in unit-level nonresponse. A variety of causes may lead to nonresponse, including refusal to participate because of concerns that confidentiality will not be maintained or political tensions causing citizens to avoid interaction with surveys that have some national government affiliation. This may be particularly important for panel-based evaluation studies since respondents need to be both enrolled in the study, and followed up on at a later point in time. At both points in time, if the data do not allow analysts to identify who, amongst those who should be in the data, are missing, the external validity of the findings is inevitably weakened. Again, although initially this may seem irrelevant to developing countries, since the latter typically have low refusal rates, it is increasingly relevant. Not only do overall response rates in these settings appear to be in decline, particularly in urban sectors, but the high mobility of many developing country populations makes attrition particularly problematic for longitudinal surveys of the type favored by evaluation research (roughly 5 percent per year in the rural Malawi Diffusion and Innovation Change Project (MDICP) survey, and 9 percent in the Mexican PROGRESA survey between 1997-99 rounds).

The next broad category of error refers to respondent accuracy. The first to be examined is nonresponse error at the item level. Even where a respondent readily agrees to participate in the survey, s/he can either refuse to answer a specific question, or choose to answer it with a "don't know" response. Comfort levels in responding to questions about income, sexual activity, fertilizer use, HIV testing, and ideal number of children, for example, will differ between people and not all will agree to provide answers to certain questions. This can even occur on questions where researchers think (or know) that the respondents could give a true answer. In fact, it may be more likely to happen with such questions, to the extent that they deal with sensitive behaviors that researchers can be sure that the respondent has not forgotten; or attitudes that the researcher knows that the respondent has (and does not have to formulate in response to the question). Although these types of item nonresponse can be treated as informative answers in their own right (Schaeffer and Thomson, 1992), researchers typically correct them during "data cleaning"—an issue which is raised again below. But in either case, since analysts often recode them to fit what they consider a less ambiguous answer category, it raises the likelihood of error.

Measurement error due to respondents refers to a related problem. It occurs when respondents provide inaccurate answers because they have forgotten a true answer or because they choose to hide it. "Response-effects" models—building mainly on work by cognitive and social psychologists (e.g., Tourangeau, Rips and Rasinski, 2000)—have described in great detail the process by which respondents work through potential answers to a given question. Whether or not this error is systematic or random will depend on the type of question being asked. Sensitive questions, for example, such as those dealing with abortion or income over the past 12 months, may lead to systematic biases among respondents, whether due to a conscious effort to underreport sensitive behaviors, or to an unconscious error stemming from cognitive processes such as "telescoping." Either way, the key point is that respondents are autonomous social actors who can choose from a variety of "satisficing" answers (Krosnick and Alwin, 1987). That is, they don't need to expend full cognitive energies on providing the best possible answer. Nor does their ans-wer need to be accurate. They can adequately fulfill their role as respondents by merely satisfying the technical requirements of the question with a plausible answer.

Measurement error due to interviewers is somewhat different, though its effects are also measured through differential response patterns. The literature on interviewer-related error in data collection is enormous and dates to the 1920s (e.g., Rice, 1929). Since Sudman and Bradburn (1974), a core distinction has been made between "role-restricted" and "role-independent" interviewer effects. "Role-restricted" effects refer to differential response patterns stemming from interviewers' different interviewing styles and are typically measured using intra-correlation coefficients. Even with a standardized questionnaire and standardized interviewer training, such effects can be significant, especially on questions that some interviewers, but not others, have difficulty asking.

Bignami-Van Assche, Reiners and Weinreb (2003), for example, show that up to a quarter of the variance in response patterns on sensitive questions stem from role-restricted effects, making it much harder on analysts to identify relations between these variables and others. "Role-independent" effects, in contrast, refer to the effect of an interviewer's social identity on respondents' answers. This covers response variation stemming from differences in interviewers' gender, age, race, ethnicity, and so on. This is a fruitful line of enquiry for developing country scholars, not only because of frequently high heterogeneity in ethnic

boundaries, and gender differentiation, but also because, as we shall see below, it is easy to build in ways to evaluate each of these in research studies (these effects are typically estimated in regression-based frameworks).

Survey organization issues cover a large range of decisions, many of which can substantially affect the level of error to be found in the TSEM.

Mode effects, the term used to describe how data are to be collected, are described first. Traditionally, developing country researchers have had little choice about what data-collection mode to use. Low levels of literacy have meant that they relied on some form of face-to-face interview during which the interviewer would read the respondent questions, and then note the respondent's answer on a paper questionnaire. They were also expected to try to ensure privacy, as demanded by the "confidentiality assurances" that have been a part of interviewers' opening, introductory statements since the 1950s. With time, a number of alternatives to this traditional arrangement have emerged. For example, over the last few years a series of experiments with computer-based, self-administered questionnaires have been fielded in developing countries, including rural areas (Mensch, Hewett and Erulkar, 2003 and Mensch et al., 2008b). Standardized data on a relatively small range of variables have also been collected with semi-structured, paper-based questionnaires (Massey, 1987), and with more open "conversational" interviewing styles. The choice between these competing methods is often driven by a combination of researchers' scientific preferences and projects' resource constraints. But that choice has considerable implications for the range of error one can expect to generate.

More specific to mode effects in developing country surveys—of the type in which much evaluation research is often conducted—are two discrete debates. The first is about respondent privacy. Briefly, confidentiality assurances emerged in developed countries. Moreover, in many developing countries, rights to personal privacy are often understood to be normatively "Western." This has a couple of implications for how data are to be collected—and therefore the types of error that researchers can expect. First, there is some question about the extent to which emphasizing confidentiality and "response anonymity" is culturally appropriate in all non-Western settings. Some have argued that rather than put people at ease, it may actually sow seeds of suspicion and mistrust (see Back and Stycos, 1959 for a classic fieldwork account, and Weinreb, 2006 for a more general description). Second, very few survey researchers collect data

on what is known as "third party presence" or "co-presence" during interviews. But a combination of household structure, family size, and general underemployment means that, in poorer developing societies, it is often difficult to "secure" a private spot for the duration of an interview. This can be seen in some unusual data from Malawi in which interviewers, though trained to insist on a private spot for the interview, were also asked to note down co-presence at three distinct points during a 1.5 hour interview. At each of the three cases, a third party presence was noted in almost 10 percent of the interviews (Weinreb and Sana, 2009a), with different people, primarily household members, coming and going all the time.

Likewise, although national populations in many developed countries tend to be linguistically homogenous, the same is not true in most developing countries. In such cases, researchers need to translate the survey instrument into multiple languages. This, too, is problematic since it is difficult to translate a questionnaire. Consequently, the more languages into which a project must be translated—unavoidable in large, nationally representative surveys in many developing countries—the greater the amount of error that can be expected. Moreover, the less confidence researchers can have that observed differences between language groups reflect behavioral differences as opposed to questionnaire-based differences (Weinreb and Sana, 2009b).

Comparability effects refer to the additional error introduced when data-collection methods, even where they are formally identical, generate different types of error. Examples include: difficulties translating a term from a master English language questionnaire into local languages; women (or men) interviewers being the appropriate choice for one setting but generating considerable resistance in another; a different propensity across groups/countries to admit to "don't know". In each of these cases, analysts need to be able to make a convincing case that group-level differences are not an artifact of comparability problems; that the data are, in other words, comparable. This is often a much harder argument to make than first appears.

Postsurvey error, the final level in the TSEM, refers to errors introduced after the data have been collected, usually during a data-editing stage that precedes analysis. Some of these errors can occur during data entry—though that is easily, though not cheaply, avoided by double-entering all information. Others occur through attempts to "fix" missing data and inconsistencies.

These are more problematic for a few reasons. First, there are no clear methodological standards about how to resolve inconsistencies (Leahey, 2008). Second, different types of imputation appear to be better suited for different types of analysis (e.g., single imputation for accurate distributions, and multiple imputation for estimating multivariate relations). Third, there is intriguing experimental evidence that inconsistencies and missing data are best resolved by fieldworkers prior to exiting the field rather than in the post-survey process (Sana and Weinreb, 2009). In all these cases, analysts need to be aware of which variables were cleaned and how.

B. Final Comments

There are certain conditions under which researchers could focus exclusively on sampling error, ignoring other sources of error. They could do this if they could be sure that:

- the survey sample truly represented the population of interest;
- nonresponse was negligible and not selective;
- respondents always answered questions accurately;
- interviewers—and other aspects of the survey administration—did not introduce additional distortions in the data;
- questions were interpreted in the same way across various population subgroups.

If all these conditions were met, all survey error would be sampling error and nonsampling error could be reasonably ignored. The reality is quite different, however. Not only are surveys complex enterprises and difficult to implement, but survey data—the raison d'être of the whole enterprise—are a product of social interaction between two autonomous actors, the interviewer and respondent. Since that interaction is embedded in a larger social context, and since each of these individuals responds both idiosyncratically and systematically to various types of triggers—as is the case in any type of social interaction— survey data inevitably bears the imprint of both the ***general context*** and the particular interaction. In other words, nonsampling error, or errors of commission can be introduced at every stage of the data-collection process. This fact is openly recognized in every data-collection handbook. But it is not always recognized by analysts, including those who are involved in setting up or analyzing evaluation data. Fortunately, recognition of these potential pitfalls makes it easier to plan evaluation surveys that are attentive to these problems and help to reduce their potential impact.

Before looking at how such plans can be incorporated into research design, it is useful to examine how social science—and evaluation research in particular—treats survey error.

Section 3. Standard approaches to error

The assertion that analysts do not take measurement error and its effects seriously enough in actual analysis is perhaps best seen by looking at standard methods for dealing with measurement error. That is our primary aim in this chapter. As above, it is not the intention of this manual to provide an exhaustive description or critique of such methods. For that, readers are directed to a variety of alternative texts, which provide an extensive review and examination of measurement error in statistical models: Bound et al. (2001) provide a rich review of recent literature, primarily in labor and health economics; Groves (1991) draws useful connections across the different disciplinary approaches for defining and analyzing errors—a problem of language which hinders collaboration and advancement in this field; and Groves (1989), essential reading as it offers an innovative approach for combining sampling and nonsampling errors. The principal aim of this section, then, is to sketch out the main elements of the problem, focusing in particular on econometric approaches, since these are predominant in contemporary evaluation studies. First, however, key terms and distinctions are introduced, as developed in classical measurement theory.

A. Bias and Variance in Measurement Error

Classical measurement theory typically distinguishes between two types of error. One is variable error, which has a mean of "0" across the sample. This type of error widens confidence intervals and may complicate the task of reliably estimating relations between variables of interest. The second type of error is bias. This latter form of error is often ignored and assumed to be, or at least treated as if it were, trivial in estimating relations among variables. It refers to a type of constant error affecting statistics throughout all implementations of a survey design. The common metric for aggregating both the bias and variance of the error terms is the mean square error (MSE). The MSE for any survey statistic is the sum of two terms: the variance and the square of the bias. The combination of these two is conceptually appealing, but in practice the MSE is rarely calculated for survey statistics. Its utility is primarily as a conceptual tool.

Underlying both of these concepts, variance and bias, is the notion that a "true" answer exists. This is important since it reminds researchers to clearly define and delineate whatever it is that they aim to measure. Now, assuming the existence of a true value, inaccurate measurement

can be defined as that which occurs when a recorded answer deviates from the "true" answer. This a "response error." Consequently, a "response effect" is something that causes a response error. For example, if X^* is the true value for a variable and X is the actually measured value for the same variable, the response error (ε) is the difference between X and X^*. Thus, for any given individual i,

$$X_i = X_*^i + \varepsilon_i \quad (1)$$

There are two main components of response error, response variance and response bias, allowing (1) to be re-identified as:

$$X_i = X_i^* + \varepsilon_i^v + \varepsilon_i^b \quad (2)$$

Response variance (ε_i^v) essentially refers to statistical 'noise,' which is created by individual or transient factors, in particular observational contexts. Take for example two forms of measurement error that arise in Figure 1 (in section 6) under respondent-accuracy issues: measurement error due to respondents and due to interviewers. A respondent may decide to lie; or one interviewer may be more liable to miscode than another. So long as a large majority of respondents doesn't lie in the same way on the same question, and so long as the subsample that the interviewer is assigned is socio-demographically equivalent to subsamples assigned to other interviewers, these particularistic sources of response error simply add more variance to the population estimate of X. They do not bias the estimate of the effect of X. In short, response variance is fundamentally an individual-level type of error. In the aggregate, and over repeated measurements of the same individual, there is an assumption that it is randomly distributed with a mean of zero (i.e., $E(\varepsilon_i^v) = 0$).

Response bias (ε_i^b), by contrast, draws on the opposite assumption. It is caused by the "essential survey conditions" (O'Muircheartaigh and Marckwardt, 1980) rather than by transient factors. For example, X may not be a completely valid indicator of X^* because it may also measure a second underlying phenomenon Y. Alternatively, the bias may be caused by field-measurement procedures such as data collection and recording methods, or type and behavior of interviewers. Response bias is most easily differentiated from response variance once the data are aggregated. The test for bias is essentially a reliability test. That is, ceteris paribus, the existence of response bias can be inferred from differences in the distribution of data across

survey conditions. For example, if male and female interviewers are randomly assigned to respondents in the same population and the mean response to a given question differs by the gender of the interviewer, then the existence of response bias can legitimately be inferred (even though, without validation data, we cannot know which group of responses is more or less accurate). In short, response bias (ε_i^b) is not zero ($E(\varepsilon_i^b) \neq 0$).

Groves (1989) offers a useful figure to link the variance and bias of measurement with the primary components of the TSEM presented in the earlier section. As apparent in Figure 3.1, many of the components of the TSEM can be associated with both increased variability and increased bias in the survey error. The same discussion in Groves (1989) also makes useful clarifications to distinguish the approaches of the different disciplines. The focus here is on distinctive features of the "standard" econometric approach, while acknowledging the existence of important exceptions.

First, it is important to note that the discussion is based on the context of a single survey project, administered in a single country at a point in time. This means that comparability effects (discussed in Section 2) are not really present. They could of course be consequential in panel data if methods, approaches, or individual responses to survey conditions change across rounds of data collection.

As noted earlier, the MSE provides an indicator of the total error produced by both the variance and bias in survey measurement. In each case, the variance and bias in error are divided into error that is due to errors made during observation, such as in the course of an interview, versus errors that are generated outside the interview itself, such as those related to sampling errors or using a poor sampling frame and obtaining bad sample coverage. The errors of non-observation include coverage, nonresponse and sampling. Nonresponse errors in this case refer to unit-level nonresponse, whereas item-level nonresponse is assumed to fall within the respondent error category below. As shown here, nonresponse (at the unit level) is associated with both variance and bias in measurement.

Respondent- and interviewer-related errors are errors of observation, and may be associated with both increased variability and biases in responses. Thus, different approaches used by interviewers—either different interviewers in the same survey or even the same interviewer across different respondents—may either lead to higher variability in responses, or to a systematic error in one direction or another. Likewise, respondent motivation may also be the

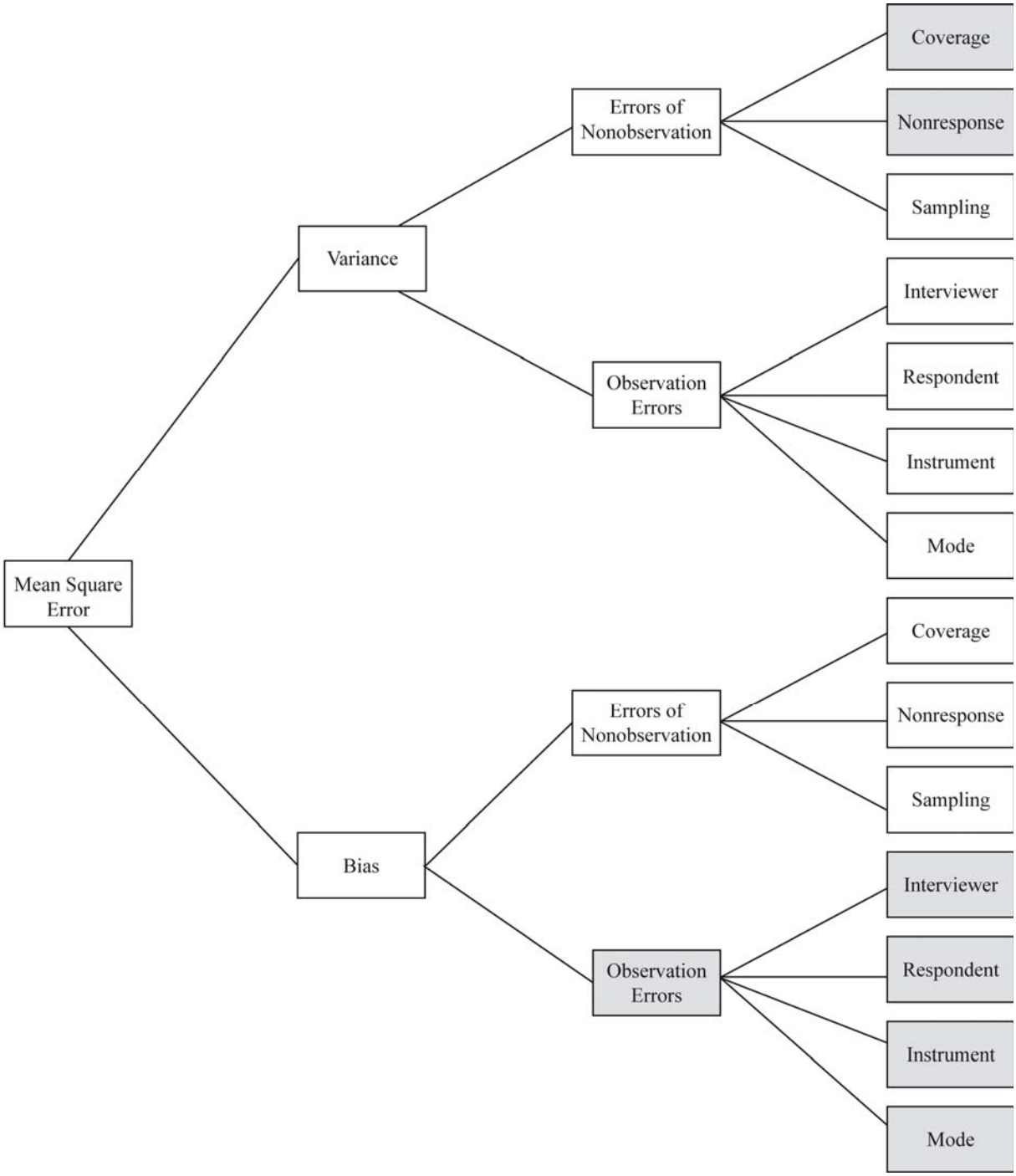
cause of random errors which lead to greater variance, or systematic errors that lead to bias. Both the survey mode and the instrument will in many individual survey efforts remain relatively constant, so that this factor does not play much of a role in this discussion. However, if a survey instrument varies over time, or even the ordering of questions is to change, it may lead to more bias or variance. A similar comment could be made about the mode of the survey.

According to Groves (1989), important differences have emerged in how different disciplines approach survey errors and which components enter into standard calculations. The econometrics approach is obviously concerned with sampling probability and how errors of non-observation affect the variance of estimators. A great deal of interest has been focused, as discussed below, on the errors-of-observation component of the variance as well, where these errors and their consequences are directly considered in the errors-in-variables model. These errors, however, are assumed to be primarily stochastic rather than systematic.

Likewise, errors in non-observation are increasingly considered under the rubric of the selection bias literature. Here, it is understood that systematic processes in the selection of cases through coverage, sampling or nonresponse may lead to biased estimators. For example, when specific subgroups of the population are missed, then the errors of non-observation create selection bias, which selection bias models attempt to treat ex post.

Finally, the shaded cubes in Figure 3.1 reflect specific components of error that traditionally receive far less attention in econometric approaches to error. These include observation errors that are associated with bias; systematically "bad" data are not handled as conveniently in this approach. Furthermore, errors of non-observation such as coverage and nonresponse that are associated with variability are also prone to less examination, although this is certainly a less accurate reflection of the current state of the econometric literature than it was when Groves first wrote this in 1989 (see Bound et al., 2001).

Figure 3.1 The Sources of Measurement Error from an Econometric Perspective



Source: Adapted from Groves (1989: p. 29).

B. Survey Nonresponse Problems

The iceberg model presented in Figure 2.2 defined sources of error in the TSEM. Nonresponse error, which includes both unit nonresponse and item nonresponse, raises important concerns for analysis in general, and evaluation in particular. In both cases, considerable progress has been made in developing methods that treat missing data (Little and Rubin, 1987) and selectivity bias (Wooldridge, 2002). Consequently, much is now known about what causes "selection bias"—the term used predominantly by economists in this field—and how it leads to problems in the statistical model. Briefly, when the unit or item nonresponse is itself not correlated with explanatory variables in the model, the problem can be ignored. In the missing data literature, this is referred to as "missing completely at random," or MCAR. When there is a correlation, various methods such as Heckman's two-stage estimator (Heckman, 1979) or preferably multiple imputations can be used (Little and Rubin, 1987). We refer readers to Wooldridge (2002) and Little and Rubin (1987) for more detailed examinations of these topics (see also Cameron and Trivedi, 2005).

Notwithstanding these advances problems remain, operating in particular through the lower panel—the bias/errors of non-observation mechanism—in Figure 3.1. An analysis of PROGRESA effects and incentives reported in section five highlights this in more detail, in relation to the extent to which survey participation is associated with possible program incentive effects. In the absence of other types of data, this type of selection bias is very difficult to deal with.

C. Mismeasurement effects on univariate statistics

Evaluation research is often concerned with univariate statistics. When measurement error arises here, the implications relate directly to the variance and bias and are quite straightforward. Where mean error is zero (E(e)=0), mismeasurement will simply generate "noisier" statistics. This is not particularly problematic where the focus of the analysis is on relations among variables—though it shouldn't be completely ignored here either. But where the focus is on the statistics of diversity, it is a problem, since noisier statistics artificially increase variance. Given that evaluation research does not usually limit itself only to the mean— on a given behavior of

interest we often ask about people at the extremes of the distribution —noisier statistics are therefore a concern.

A more direct effect on evaluation is to be found where mismeasurement expresses itself as bias ($E(e)\neq 0$). In this case, all univariate statistics are affected, thereby altering any evaluation statistics that are based on these data. For example, if everyone in a sample reports income levels about 10 percent below the "true" levels, then the sample estimate will be 10 percent too low. More worrying, new or innovative behaviors—frequently the focus of evaluation research—are often associated with a relatively small number of individuals. As a rule, those new behaviors tend to be underreported (as they challenge existing norms and interests in the community). This underreporting bias can lower sample estimates dramatically. For example, where estimated sample size is based on a 10 percent prevalence of some behavior, but that behavior is reported by only 3 percent of the sample, then the data are, to say the least, problematic. As is shown below, this may not affect the estimated relationship between income and some other variable (assuming no selection bias in who reports correctly and who does not, the intercept will be affected, but not the coefficient). However, even if researchers' aim is merely to track income over time, the bias will remain. Each and every estimate of income will be biased.

Perhaps because univariate statistics are so simple to estimate, this point is often ignored in discussions of measurement error. But it is an important one, particularly in evaluation studies in developing countries. The main reason is that many such studies are interested in time trends in new behaviors, many of which are sensitive, or even stigmatized. But over time, as people become more exposed to those behaviors, or to talk about them, sensitivity and stigma—the primary drivers of social desirability biases—may fade. One notable example of this in evaluation research has to with the fertility transition in East Africa, particularly Kenya. Researchers largely missed the early stages of the transition, not because data weren't being collected—on the contrary, multiple surveys were fielded— but because "early adopters" no longer felt themselves to be absolute behavioral outliers, making them more willing to admit to using contraception to survey interviewers. But this change in respondents' reticence to report contraceptive use—expressing itself in different levels of measurement error across time—means that the univariate trend in contraceptive use exaggerates the steepness or suddenness of the widespread adoption of family planning (see Robinson, 1992). That clearly has implications for how family planning programs should be set up.

Similar problems can easily be envisioned about many other behavioral innovations that are the subject of intervention and subsequent evaluation. In the realm of public health, one can think of condom use, frequency of HIV testing, attitudes to women's autonomy, extra-marital sexual partners (especially as reported by women), drug use, and sexual abuse. In the realm of economic behavior one can think of anything related to participation in grey or black markets, to the receipt of remittances, and—especially in poorer developing countries—the need for support from external sources. The contextual factors that drive some of these biases are addressed in more detail in Section four.

D. Mismeasurement effects on bivariate and multivariate statistics

The effects of mismeasurement on bivariate and multivariate statistics are standard econometrics territory. Econometrics textbooks provide clear guidance about measurement error, the basic elements of which are relatively familiar to most evaluation researchers. Specifically, measurement error in these texts is treated in an ordinary least squares (OLS) framework, that is, with a continuous dependent variable and under the further restrictions of OLS. The following discussion follows closely the development in Wooldridge (2002), which offers a particularly clear exposition of the consequences of measurement error on coefficients in the OLS model. It first deals with a case where measurement error affects the dependent variable—the simpler case—and then where error affects an explanatory variable.

i. Measurement Error in the Dependent Variable

The first case occurs when measurement error exists in the dependent variable. In these cases, researchers seek to explain a variable, y^*, such as annual household expenditures. In this case, the regression would take the form,

$$y^* = b_0 + b_1 x_1 + \mathsf{K} + b_n x_n + u \quad (3)$$

We assume the model is consistent with standard assumptions for OLS, but since y^* is not known, we make do with a measure of reported expenditures, y, collected in the course of a survey. The gap between y^* and y is the measurement error in our dependent variable (e_0). It can stem from any one, or combination, of the range of factors discussed in the last section. Thus,

$$e_0 = y - y^* \quad (4)$$

which can be incorporated into (3) to show how the actual measured dependent variable, y, is estimated:

$$y = b_0 + b_1 x_1 + b_n x_n + \mu + e_0 \quad (5)$$

The new estimated model in (5) illustrates that the error term is now altered and includes both the original term as well as the measurement error in the dependent variable, e. In actuality, the model is estimated as usual since it ignores the possibility of measurement error in the independent variables. The estimated coefficients, b, are consistent in the OLS model if E(μ)=0. If E(e)=0, then the measurement error has no impact on any of the coefficient estimates. But if E(e)≠0, the bias in e will produce bias solely in the intercept, b_0. The remaining coefficients in (5) are unbiased and consistent. The variance estimate of the combined error terms is larger than the variance of the original error alone. This larger error variance will affect the significance of our statistical tests.

In conclusion, measurement error in the dependent variable, assuming it is unassociated with any of the independent variables, raises the error variance in the estimated model, but OLS maintains its appealing properties. As Carroll, Ruppert and Stefanski (2006: p. 341) note, "(a)ll tests, confidence intervals, etc. are perfectly valid. They are simply less powerful." Notably, the case where nonclassical measurement error appears is rarely discussed in this literature, but it is clear that it may lead to biased and inconsistent estimates. It should certainly be a concern when attempts are made to estimate important but error-prone measures such as individual levels of current health status or total household income over the past month or year. In such cases, the potential for a broad range of covariates to be correlated with measurement error in the dependent variables is apparent.

ii. When error is in the explanatory variables

The consequences of measurement error when it occurs in one or more of the explanatory variables is of great concern because it has the potential to both bias the estimates of the coefficients as well as increase the variance of those estimates. It is instructive here to present a version of the traditional *errors-in-variables* example, in the context of a simple bivariate

regression model. The dependent variable in the model is assumed to be measured without error, while the independent variable, x^*, is measured with some degree of error. In the first place, we assume that the model agrees with the basic regression assumptions so that OLS would produce unbiased and consistent estimators of the coefficients b_0 and b_1.

$$y = b_0 + b_1 x_1 + \mu \quad (6)$$

However, the explanatory variable, x^*, is not observed and we use the observed values of x in its place. This change is problematic because it makes the explanatory variable stochastic—a violation of the assumptions of the classical linear regression model. Suppose we are interested in estimating the effect of annual earned family income on child years of schooling, then y may represent years of schooling and x^* is the real measure of income. In reality, x^* is unobserved and we generally rely on household reported annual income, x. The measurement error in the population is then simply,

$$e_1 = x_1 - x_1^* \quad (7)$$

We can rewrite the model to be estimated in terms of the observed explanatory variable through substitution into (6) so that,

$$y = b_0 + b_1 x_1 + \left(\mu - b_1 e_1\right) \quad (8)$$

The main assumption that is made in the *errors in variables* model is that the measurement error in the explanatory variable is uncorrelated with the unobserved explanatory variable x_1^*. That is, we assume that cov(x_1^*,e)=0. This is a necessary assumption for the *errors-in-variables* model but one that is well known to be overly restrictive in reality (Bound et al., 2001). It is definitely not consistent with a considerable body of methodological literature in data collection. In particular, if salient characteristics of the respondents or of the interviewers introduce measurement error—for example, increasing accuracy by education or income of respondents or lower accuracy for interviewers from other cultural backgrounds—then this error may itself be correlated with a wide variety of explanatory variables routinely incorporated in evaluation models.

Given (7) and the lack of correlation between the unobserved value of x and the measurement error, the measurement error must be correlated with the observed explanatory variable, x_1. It can be shown that since $x_1 = x_1^* + e_1$ and the two components on the right hand side of (7) are uncorrelated, x_1 and e_1 must be correlated. Furthermore, the covariance of x_1 and e_1 is equal to the variance of the measurement error, σ_e^2. The positive covariance between the observed measure of the explanatory variable, x, and the measurement error, raises a red flag. Recall that the estimated model in (8) shows that the composite error term, $u - b_1 e_1$. This composite error term must not be unassociated with the observed measure x. It can be shown however that,

$$\operatorname{cov}\left(x_1, u - b_1 e_1 = -b_1 Cov\left(x_1, e_1\right) = -b_1 \sigma_e^2\right) \quad (9)$$

The negative covariance indicates that OLS regression in (8) will lead to a biased and inconsistent estimator for the b_1 coefficient. Using asymptotic properties, it can be shown that at the probability limit, the estimated value of the estimator is,

$$p\lim\left(\hat{b}_1\right) = b_1 + \frac{Cov\left(x_1, u - b_1 e_1\right)}{Var\left(x_1\right)} = b_1\left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) \quad (10)$$

This well known result of the errors-in-variables model highlights the impact of measurement error in the explanatory variable on the estimated b_1 coefficient. Because the coefficient is multiplied by a ratio which is always less than or equal to one, b_1 is always biased towards 0. This effect is labeled attenuation bias. It is also important to note that the extent of the bias is recognizable from the relative magnitude of the variance term in the numerator relative to the variance of the denominator, which is in fact equal to the variance of the observed variable, x_1. This result provides an informal gauge to estimate the potential attenuation bias created by measurement error, which may be small if the variation in x^* is much larger than the variation in e.

The unambiguous prediction of attenuation bias in the coefficient is a clear and compelling result. Recalling the earlier attenuation bias caused by measurement error in reported family income would lead to an overly small estimate of the effect of family income on years of

child education. However, the clear results of this model of attenuation bias would appear to suggest that evaluation researchers and other analysts can overcome the limitations of data quality by acknowledging the bias in the coefficient estimates, which is a priori known. Nevertheless, two critical caveats need to be dealt with before this conclusion can legitimately be accepted.

The first refers to the added complications that are created by including additional explanatory variables into the model and estimating a multivariate OLS model. This is clearly a much more realistic scenario in most evaluation research than the simpler bivariate model. In this case, if we maintain the above errors-in-variable assumption that the measurement error and the observed value of x are correlated, then our estimate of b for the variable that is measured with error remains biased and inconsistent and the same attenuation bias may hold. However, it is also the case that all other variables in the model may be biased and inconsistent, even if they are not measured with error. Furthermore, the direction of their bias may be very difficult to determine except with additional assumptions (Bound et al., 2001).

The second caveat involves the assumptions that are necessary for the errors-in-variable result to hold, even in the bivariate case. Although these assumptions are frequently invoked, they may not be tenable in many cases. Many of the concerns cited earlier about response effects raise potential problems with the errors-in-variable assumption in that the explanatory variables are uncorrelated with the error in measurement. The literature on response effects is replete with evidence showing how social and cultural characteristics alter response error due to factors associated with the respondent, the interviewer or both. Each of these associations may contradict the assumptions of the errors-in-variable model.

Three other points should be made. The first is that extensive theoretical and empirical studies have been done to expand the utility of the assumptions that are required in the OLS model (Black, Berger and Scott, 2000; Meijer and Wansbeek, 2000). Instrumental variable methods have also been used to try and overcome measurement error restrictions, although they often demand additional restrictive assumptions (Antman and McKenzie, 2007; Hu and Schennach, 2008).

A second point is that nonlinear models raise a host of other concerns that appear to require restrictions that are at least as strong as those required for the OLS model (Bound et al., 2001), though there are important advances being made in this area of research on measurement

error as well (Hausman, 2001; Schennach, 2004; Hu and Schennach, 2008). Because much evaluation research falls into this pool —with many outcomes of interest dichotomous or categorical in nature— even the basic findings of the errors-in-variable model may not hold.

A third point is that fixed-effects models, which offer so many advantages for evaluation research and are an underlying motivator for the increasing popularity of panel surveys—may themselves exacerbate the potential for measurement error to affect findings. Essentially, the concern with measurement error can grow since removing the fixed effects eliminates alternative sources of error, leaving measurement error a larger fraction of the remaining variation (Wooldridge, 2002; Griliches and Hausman, 1986). Thus, consider the case where efforts to overcome traditional problems of unobserved heterogeneity in the estimation of the impact of family income on child schooling success, which is measured with data on grades, lead to a fixed-effects specification. For this purpose, panel data are used, providing annual measures of family income and annual data on child grades. However, the same fixed-effects estimator, which enables one to control for unchanging and unmeasured differences between households, also increases the proportion of the variation in child schooling that is caused by measurement error in the explanatory variable.

E. Conclusion

Our central point in this chapter has been to suggest that whether a researcher's goal is to reliably estimate the frequency of a given phenomenon or estimate relations in multivariate frameworks, measurement error—which inevitably exists to some degree in both dependent and independent variables—poses a serious problem. It can mislead researchers interested in identifying a univariate trend over time. It can reduce the chances of correctly identifying a relationship between two variables that actually exists, or generate biased estimates of that relationship. In fact, this is all the more likely in the case of nonlinear models of the type that are often used to assess evaluation programs (given categorical or dichotomous outcomes such as attendance at a health clinic, school attendance, contraceptive use). Finally, this chapter has also shown that measurement error may be even more problematic for panel data—of the type favored by evaluation research—since here, in addition to idiosyncratic error in the cross-section, analysts must deal with shifting error and sources of error across rounds (e.g., changes in the acceptability

of a given intervention; different types of interviewers or questionnaires; or different attitudes to the research stemming from changing political context).

Observed changes in the dependent or independent variables at the start or end of a panel can stem from shifting levels of measurement error in either round. (A particularly salient example of this would be program effects if programs themselves caused changes in the quality of data collected—this issue is examined in a subsequent section.) Ultimately, therefore, while econometricians have been successful at generalizing the conditions under which classical measurement error generates consistent findings, in many cases the use of these traditional assumptions reflects "convenience rather than conviction" (Bound et al., 2001). There is no magic wand. The problems reflect difficulties that are deeply entrenched in data on human behavior, where those data are themselves the product of human behavior and interaction.

Given the difficulties posed by measurement error, the inadequate fixes, and the problematic assumptions, it seems reasonable to suggest that the best solution to the problem of measurement error is to employ a complementary strategy, the first stage of which seeks to reduce measurement error as much as possible during the earliest stage of research: the data-collection process. The following section lays out a step-by-step series of recommendations on how to go about that.

Section 4. A Brief Guide to Reducing Error at Source

Reducing error at source builds on two general principles. The first is primarily organizational. Both large-scale surveys and smaller research studies involve multiple players, including those who have no formal role in the project. Each of these players has different goals and interests. This inevitably affects the extent to which they either care about mismeasurement or are able to affect it.

The second principle is primarily epistemological. Since levels of error are unknown and cannot be known a priori, researchers should incorporate elements into the study design that will enable them to assess the level of error prior to analysis, and evaluate its effects on their data. Without such elements, the degree of confidence in their results will inevitably be weaker.

This section expands on each of these principles, in each case describing specific data-collection recommendations that follow on from them. Note that the specific relevance of each to particular types of evaluation varies, primarily based on how large the evaluation is, the extent to which it is conducted in collaboration with the government, and the extent to which researchers can choose—or veto—elements of the study design. Also note that for the sake of brevity and readability—boiling down fieldwork recommendations to a single chapter is no small task—recommendations are made almost wholly within a narrative framework. That is, no models or formal analyses are presented here; they merely are referenced where they exist, awaiting the final chapter which contains an empirical focus on a key emerging issue in least-developed country (LDC) fieldwork: gifting.

A. Survey administration

Surveys, including survey-based evaluation studies, are organizationally complex undertakings conducted within a set of financial and time constraints. The larger they are, the more complex the organizational structure becomes. Before directly addressing how to reduce error in such an enterprise, it is useful to review the basic steps involved in fielding a project.

Choices need to be made at multiple stages. In rough chronological order—note that those discussed in more detail later are marked with an * below—researchers need to:

1. Choose a country, a site (or sites) in that country, and local collaborators. These decisions will affect how involved they are in day-to-day running of the project.

2. Find and employ a local survey director and field supervisors. Sometimes researchers can use, or may be compelled to use, personnel from already existing teams, such as those maintained by national statistical offices or survey research groups.
3. Establish access to cash, for example, by opening a bank account into which project funds can be deposited—someone needs to pay for hotels, supplies, salaries, transport, hardware, and so on. Note that accomplishing this task can be more difficult than it sounds given a researcher's temporary status in the country.
4. Arrange for transport, which may include renting or purchasing vehicles, and hiring dedicated drivers. Researchers need to assess the type of transport that they will need given the place and season (how are the roads? Is it the rainy season?).
5. Arrange sleeping and related administrative accommodation, like temporary office space, in or as close as possible to the sampled sites. This can also mean setting up an independent "off-grid" site using, for example, solar panels or car batteries to power laptops.
6. Begin prefield work preparations such as training of field supervisors, preparing for sampling, and questionnaire translation.
7. *Have field supervisors conduct a field test with the initial version of the translated questionnaire.
8. *Based on the pretest results, finalize and translate the questionnaire.
9. *Hire and train interviewers. Researchers should do this with the help of the field supervisors. Since the supervisors will be working with these people and responsible for them, they have an interest in hiring good people and training them well.
10. *Pay respects to local leaders.
11. Set up a logging system for the sample, so that the status of each sampled individual or household—completed interview, first contact, refused, not yet contacted, etc—can be known at a glance, and so that no one is omitted and no one interviewed twice.
12. Finalize arrangements for data entry, whether done in the field under the direct supervision of the project (including CASI interviews where the data are entered by the respondent), or sent to a specialist data-entry company.

Juggling these tasks is complicated at the best of times. It tends to be even more complicated given the inevitable tensions that arise between competing interests. Those tensions are generally overlooked in the methodological literature, though they are the subject of much fieldwork lore, and certainly leave their mark on data quality. Consequently, those are the subject of the next subsection.

B. The players and the context

Given the complexity of surveys, including survey-based evaluation studies, a range of people—"central players"—are formally involved in the study. There are also people who are not formally involved in the study but who are affected by it, who may have an interest in it, and can bring influence to bear on the actual participants. These are "potential spoilers."

The specific characters in each of the two categories—those who are officially involved in the study and those who are not—are listed in Table 4.1. It is important to note that although there is no formal methodological literature on the effects of all these players on data collection, they figure prominently in ethnographic accounts and fieldwork reports in general. Thus, the list in Table 4.1, and subsequent discussion, draws on a wide array of studies, including the early methodological literature on LDC survey research (a sample of which was given at the end of the introduction), reviews of World Fertility Survey field practice (e.g., various chapters in Cleland and Scott, 1987), critiques of standard LDC survey practice (e.g., Stone and Campbell, 1984), methodological literature in anthropology (e.g., Agar, 1980; Barley, 1983; Bernard, 2000), as well as informal conversations with researchers who have organized and implemented LDC surveys. The overall message from this literature is that distinguishing between these types of players and, more generally, becoming familiar with the political contours in which an intervention and evaluation is conducted, helps us understand how data come to take the form that they do.

Table 4.1 Typical LDC survey players, by formal involvement in the project

Directly and formally involved in the project	
YES	NO
Program designer & funder Lead investigator Local elite collaborator Field director, supervisors, data editors Interviewers and ancillary staff (e.g., data entry) Respondents	Local political and administrative leaders Local religious and business elites Respondents' family members Local nonrespondents

Two final points about these players: all of them, whether formally involved in the study or not, are embedded in a political context that, in many LDCs, includes a somewhat ambivalent relationship to government. They are also embedded in an economic context that, likewise, often includes substantial unemployment or underemployment, and limited opportunities for resource accumulation. These contextual factors also affect local interpretation of the study and its goals, and local responses to the study, including survey responses.

Central players: People formally involved in the study

The critical factor to take into account is that each of these players—beginning with those in the left-hand column of Table 4.1—has a different interest in the research. Program designers and investigators tend to be highly skilled, highly paid specialists who are primarily interested in the study's results. They also tend to be involved in other projects—writing-up completed work, applying for grants for future work, teaching, and so on. All of this means that they have much less interest in spending months overseeing fieldwork than in being presented with a data file so that they can generate evaluation or research results.

Local elite collaborators have a somewhat similar profile. They, too, may also be interested in the results and will expect to be a co-author on any final paper(s) that use the data. But because they are often working on other projects they are also unlikely to spend a considerable amount of time in the field. The best LDC-based researchers are famously overextended; they are also often formally associated with local research institutes that are

financially dependent on project overheads and that push their affiliates into these types of collaborations.

Inevitably this means that field staff—at least after an initial training period—take full charge of data collection. And herein lies the problem, at least in many studies. Since field staff are rarely, if ever, involved in data analysis and publication, their responsibility ends as soon as they hand off the data to the lead investigators. Their primary goal is therefore to do a good enough job to guarantee themselves the prospect of more such work in the future—in many developing countries, interviewers often earn five to ten times as much per day as equivalently educated public sector employees (e.g., teachers, nurses). Moreover, they know that "good enough" tends to be judged in only two ways: the pace of data collection (it must stay within budget); and the overall level of coverage and nonresponse on completed questionnaires (i.e., anything related to respondent-selection issues and nonresponse, as appears in Figure 2.1). This is the case whether one is talking about senior field staff or more junior ones like interviewers. In each case, fieldwork is a job. It puts money in fieldworkers' pockets, allows them to remit to dependent family members, to build up capital for a post-fieldwork business, etc. One might hope, of course, that interviewer-selection procedures are good enough that most interviewers will have sufficient intrinsic motivation to do their job well. But even if that is the case, it doesn't mean then they care about mismeasurement in the way that analysts do – or should.

Finally, the last in the list of formally affiliated players: the respondent. They probably care even less about the research than any of the other players. They are also much less likely to understand the difference between different types of research (e.g., evaluating a theory- driven hypothesis for the sake of science versus a baseline study that will precede some development-related intervention). Given the possible lack of trust in government and other nonlocal institutions, they may also be less likely to even believe that the project's stated goals are its actual goals. In either event, and as noted above, a respondent's job is simple: it is to provide "satisficing" answers to an interviewer's questions (Krosnick and Alwin, 1987). These can be true answers but they don't have to be. And again, whether or not they are more or less true is a function of the respondent's own motivation, and the effort that they are willing to invest in responding (Tourangeau et al., 2000; Schaeffer and Presser, 2003).

The different roles and goals of each of these formally affiliated groups of players have clear implications for data quality. Those with the most substantial interests in minimizing

error—the program designers and analysts—are rarely on hand to minimize it. Those who are on hand are temporary staff, "alienated"—in classical Marxian terms—from the product of their labors. They know that they will receive no bonus for ensuring high quality data beyond the easy-to-spot issues of coverage and nonresponse. The same is true of respondents: for them, the decision to provide a plausible answer that involves little effort, as opposed to a truer one that requires more effort—or even some psychological cost—will not be rewarded or punished; one answer is as good as the other from their point of view.

Potential spoilers: People <u>not</u> formally involved in the study

The list of involved players extends to those who are not formally involved in the study. The right-hand column of Table 4.1 lists these types of people. Although they are rarely discussed in methodological discussions, they can all too easily affect data. Here they are referred to as potential spoilers.

Topping the list are different types of local leaders. It is important to distinguish conceptually between political and administrative leaders, who are formally associated with the state, and religious and business leaders, who are at least nominally independent. But in general, a research project should handle them similarly. Specifically, researchers should aim to forge cordial ties with all such leaders, but also to maintain as much actual independence as possible, and certainly <u>an image of institutional independence</u> in the eyes of respondents. Thus, project representatives should pay respects to local leaders but not become, or be seen to become, affiliated with any particular leader or faction.

There are a couple of ways in which this principle should be enacted in practice. First, even if a project has received the blessing of key figures in the central government, lead investigators and senior field staff need to visit with local political and administrative leaders, to pay respects, and to be seen to pay their respects. This is true where the central government is liked—not paying sufficient respect to a local leader in this case can be seen to belittle him. But it is also true where the central government is disliked, since in such cases, research projects do not want to be seen as agents of the government, or to be imposing yet another "central government" decision without local consultation. In brief, the last thing that projects want is for local leaders to urge residents—openly or covertly, from a local government office or a church pulpit—to openly refuse to participate, or to claim that they are "too busy" to participate.

An extension of this concern covers the temptation that researchers often feel to collaborate with local authority structures. Although this type of collaboration can smooth all sorts of local processes, it should be avoided where possible. The reason is that in every setting (including closed political systems) there is competition for political power at the local level, including competition that expresses itself through mobilized ethnic, clan, and religious identities. If researchers are seen to collaborate too closely with a local leader from one such faction, they may unwittingly generate nonresponse or response bias in samples associated with opposition factions. Again, the guiding principle here should be cordial relations but institutional independence.

Likewise, if a local administrative leader is kind enough to offer use of certain facilities—an office in his building, use of public vehicles at subsidized cost—researchers need to be wary. They need to think about whether the public perception of their affiliation with that particular administration is going to generate response bias. For example, will reported health behavior be different if interviewers are driven into a village or neighborhood in a large 4 x 4 vehicle with a "Ministry of Health" emblem on the door? Will reported behavior of any sort be affected if it is widely known that the project's base-camp is in an annex of the local city council building? Will that make people more or less likely to trust interviewers—or any other representative of the project—with sensitive personal information?

Even if one must largely turn down offers of assistance (where one can) it is important to forge connections. The easiest way to do this is to give local political and business leaders—there are often close connections between these groups —something of a stake in the research. This does not mean opportunities to co-write/co-publish results. Rather, it means recognizing that local leaders often see the project as an opportunity to generate some capital inflows into the area. One of the best ways to do this is to hire qualified people locally: there is almost never a shortage of qualified and able people in LDCs, and as much as unemployment and underemployment tends to be bad in urban areas, it is even worse in rural. And they can be put to good use in any number of ways. They can match questionnaires to local dialects, map research sites, introduce interviewers and supervisors to targeted households, or serve as key informants about a range of local issues (much of which can be standardized and coded, allowing for the identification of community-level differences between different research sites). They can also serve as interviewers themselves—an issue addressed below. For now, suffice to say that using

local interviewers is particularly useful since, although it means replicating interviewer training in every large research area, it improves response rates, appears to improve data quality (in LDCs), and cuts project costs (they don't need hotel accommodation). Overall, it allows the project's leaders to impress on local leaders how much the project is contributing to the local economy.

A similar point can be made in relation to local business leaders. Research projects, in particular those that aim to maintain a relationship with an area over time—as is the case with many panel evaluation studies—should purchase at least some of their supplies at local businesses. Where there are multiple local suppliers, they should take care to spread the purchasing among them. This can include simple things like pens, paper, and food. Or it can include more expensive items such as "gifts" for respondents—to which we will return below. The point is to maintain some flow of resources into local businesses since that is one of the most effective ways to give local leaders a stake in the research. And that stake can be very useful if the project runs into problems in other ways, since it means that the lead investigator or field director has some form of leverage.

Respondents' family members can play a somewhat different spoiling role. In fact, this is the one type of potential spoiler that has made it into the mainstream methodological literature. As mentioned in section two, notwithstanding interviewing norms, there is considerable third-party presence in many interview situations (though it is seldom acknowledged). Some of this co-presence stems from a curiosity factor: a stranger is sitting in the living room asking questions of a household member. Some of it also stems from the fact that interviewing certain types of respondents can be understood as upsetting familial authority structures: e.g., a stranger is sitting in the living room asking questions of the young wife of one of the patriarch's sons rather than her father-in-law, mother-in-law, husband, etc. Either way there is co-presence, and it is not random.

There are two things that can be done to minimize such effects. First, survey interviewers should temporarily stop the interview and contact the survey supervisor. The latter should arrange for an additional interviewer to come and simultaneously interview the meddlesome third party at the same time as the sampled household member. These bonus data can either be immediately discarded or, more usefully, they can be entered with other data, flagged as nonsample, and then used to estimate within-household consistency checks.

Second, while designing the survey instrument, researchers should include a question-to-the-interviewer on the same page as sensitive questions. It will require interviewers to quickly note whether or not there is a third-party presence and, if so, some key characteristics (e.g., age(s), gender, relationship to respondent, and whether they are actively interfering or just listening). Consider this a part of the metadata (see also Grosh and Muñoz, 2000). These variables can then be used (within a standard regression framework) to explore patterns of response variation.

The last category of potential spoilers consists of nonrespondents. Typically, a combination of clear eligibility criteria plus the general desire to generate a geographically diverse sample means that research studies do not enroll all local residents, or representatives of every household, in the study sample. Only a fraction is selected. The problem here is that some individuals not selected for the study may want to have been selected. And they might not believe—or understand—fieldworkers' explanations that people were selected at random, since outside the world of experimental science, the benefits of random assignation are somewhat abstruse and often at odds with a perception of fairness.

Note that nonrespondents' desire to be selected for a study may be an issue where the project openly "gifts" respondents—that is, provides them with some compensation or token of appreciation for their participation—since nonrespondents are likely also to want that gift. But even where there is no gifting, people may also want to be selected. In many settings, interviews may be one of the few occasions in which people are able to interact with complete strangers. Some people are attracted by the prospect of an interview with a stranger. Others might want to be interviewed because they want to have their voice heard, or to have an excuse to take a break from the tedium of everyday tasks.

This desire to be interviewed is important for measurement error because it intersects with issues related to the respondent's identity. In many LDC settings interviewers have no idea what the person they are supposed to interview looks like. They show up to a randomly selected household looking for someone who fits certain eligibility criteria. Or else they show up with a specific name that a prior team member has already determined is the eligible respondent. This is the flipside of the nonresponse issues that so concern researchers. Just as there are people who want to avoid interviews, there are people who really want to be interviewed. Thus, fieldwork lore includes stories about nonrespondents who step in and claim to be a person who fits the

criteria, or to be the specific target person. They even can make this claim within earshot of others in the household, who go along with it.

One other reason for making this claim, apart from the desire to be interviewed or to receive the gift intended for the target respondent, is simply to have fun tricking the gullible outsider to whom they owe nothing. In other words, tricking the interviewer, particularly an outsider interviewer, can be thought of as a specific type of "sucker bias" (for other examples see Agar, 1980; Anderson, 1986; Barley, 1983). Whatever the motive, there is a convergence here of respondents' and interviewers' incentives. Respondents have a little fun. And interviewers get to complete a questionnaire without having to chase up the missing individual.

In fact, no one would even know that there was a problem unless a follow-up round of data collection was conducted in which the right person was found and interviewed "again," this time giving information that was completely inconsistent with that under his/her name from the first round. The increasing popularity of panel studies means that this phenomenon—and effect on data—is bound to increase. And based on past panel studies, the scale of the problem seems worrisome. For example, early rounds of the MDICP, which to its credit is more open about data-quality issues than perhaps any other contemporary social science project in LDCs, show only 66 percent consistency on sociodemographic identifiers (e.g., age, number of children) across rounds (own calculation). This is similar to results on the World Food Summit (WFS) response-reliability project (O'Muircheartaigh and Marckwardt, 1980; O'Muircheartaigh, 1982). And although this inconsistency is usually interpreted as simple response error on the part of the same respondent, it could equally stem from the fact that different people have claimed to be the respondent.

The most effective way around this problem is not to request an ID from respondents. Not only are there settings in LDCs in which only adults have formal ID, but the very act of requesting it can, in certain settings, make people suspicious (since only distrusted authorities and authority figures make such requests), or insert an element of mistrust (since it means that we don't trust who they say that they are). Instead, it is best to do the following:

1. Generate a project-specific ID: collect some visual or biomarker measure of a respondent's identity at the first round of data collection then verify it on all subsequent rounds. For example, a photo can be taken at the initial round and interviewing in all subsequent rounds made conditional on finding that person (in many LDC settings,

2. Increase estimated sample size by, say, 5 percent, in order to account for the extra loss-to-follow-up that stems from having got the wrong person in the first wave. Figure 4.2 summarized the steps that can be taken to reduce spoiling effects.

Table 4.2 Summary of steps that one can take to reduce spoiling effects by people who are formally uninvolved in the research study

Before anything else • Collect basic information about major fractures and factions in the community, in particular those related to ethnicity, clan, or religious affiliation
Local political leaders • Visit them and describe the goals of the study • Offer to have them or their spouse interviewed (have it done by a survey supervisor) • Employ at least a few locals in every site, or many if possible (e.g., use them as interviewers)
Local business elite • Purchase supplies at local businesses, spreading the purchasing among multiple local suppliers where possible
Local religious leaders • Give interviewers a day off – and note that this may be a Friday or Saturday for certain groups (e.g., respectively, Muslims and Jews and Seventh Day Adventists) • If planning to go to a church service, find out how much tension there is between your preferred church and others in the area (don't go if tension is high)
Meddlesome family members • Arrange for meddlesome 3rd parties to be interviewed at the same time as the sampled household member • Include question-to-interviewer around sensitive questions on the questionnaire about 3rd part presence
Nonrespondents • Photograph respondents in the first round of data collection and verify their ID in follow-up rounds – though beware IRB constraints • Remind nonrespondents that the project is providing wider benefits to the community

C. More narrowly technical guidelines

The last subsection showed how key contextual factors can affect data in highly measurable ways—for example, by reducing participation, through third-party presence, or through false claims to be the respondent. The current subsection describes more narrowly technical guidelines

about how to reduce measurement error or, equally important, the steps that we can take that will allow us to evaluate the level of error in our data. Unlike the discussion thus far, these issues are tackled in a relatively chronological fashion.

Note that space constraints mean that not every issue or stage of fieldwork is covered or can be covered. For example, nothing is said below about how to select survey supervisors, set up logging systems, organize data entry or clean data. Rather, the issues covered are either the most important sources of error, or the most frequent subject of debate amongst fieldworkers.

i. Sampling

It's not always possible but it is absolutely best to construct your own sampling frame and do your own sampling. If you inherit someone else's sample of fifty enumeration areas— for example, a census was conducted the three years previously, your local collaborator has access to those records, and claims that there has been no substantial in-or-out migration—check it. For example, send field teams to five of those areas to ensure that the sample looks right (they can map all households, identify those with eligible respondents, and see if it looks spatially random). If the budget can support it, contract a local pilot to take some ariel photos of sample sites. If not, use Google Earth. In either case, print a map and have field teams mark all listed households. It will give you an idea of how much noncoverage bias there is in the existing sample lists.

Likewise, when a local headman or chief kindly volunteers to compile a list of all local residents for your sampling frame, politely accept but then before fieldwork begins send a field team to find those households and then, as above, compare them spatially to some maps. In both these cases, the last thing that you want is to find out later—e.g., after data collection—that the local headman forgot to mention a cluster of households on the other side of the village that he was on bad terms with; or that the older sampling frame, compiled five years ago, didn't include a new area of migrants from a different region, who had arrived and set up shop three years ago.

ii. Pretest

Pretests—small trials of a draft questionnaire administered close, but not too close, to a sampled area—are time-consuming. But they are also important. This is particularly the case where a project is using new questions or where supervisors have differences of opinion about whether a

given question can be asked (without offending too many people too much), how it should be coded or translated, whether and when a gift should be given, and if so, what it should be.

The key reason that pretests are important is that neither foreign researchers nor local elite collaborators know everything. Even if the local elite collaborator once knew the area well, s/he may have lost touch with changing conditions on the ground (especially in rural areas, given that elites are typically long-term urban residents). Thus, s/he may not know that a question that would have completely alienated people twenty years ago may be completely acceptable today.

Additionally, those involved in running data collection are naturally conservative insofar as they will usually find it easier to redo something familiar and proven to work than to try something new. Remember, the incentives associated with field staff—discussed in subsection B—are different from those of lead researchers: the latter profit from innovation; the former do not.

Once the pretests have been conducted, researchers need to assess the quality of data. This is done in two ways: seeing how responses vary on central variables of interest; and debriefing the field team to see how they think the questions performed. If you hired good people and trained them well, their impressions should be reliable. Based on the assessment, researchers can then make necessary corrections to the questionnaire or to the fieldwork protocol, with every confidence that those corrections are based on empirical indicators rather than researchers' instincts, or ungrounded methodological norms.

iii. Questionnaires

Questionnaire design and modes of data collection are amply covered in many excellent texts and manuals. Likewise, the special characteristics of many developing country populations—in particular limited literacy and infrastructure—place natural limits on researchers' choices of questionnaire mode (respectively, no type of self-administered questionnaires or no phone interviewing outside selective subsamples). The discussion here, therefore, is based on two assumptions. The first is that data are to be collected in a face-to-face interview using either paper or PDA/laptop format. The second is that the interviewing style will be relatively traditional. That is, it will be anchored by a structured (and translated) questionnaire, even if it allows for what is often referred to as a more "conversational" style. The latter refers to interviewers' ability to temporarily move away from the questionnaire-based script. For

example, they are able to probe or rephrase a misunderstood question in ways that are consistent with the question's intent. This style of interviewing is the current standard in survey-based data collection in more developed countries (MDCs), (Suchman and Jordan, 1990; Schober and Conrad, 1997; Maynard and Schaeffer, 2002; Schaeffer and Presser, 2003). Its widespread use in LDCs is one of the likely explanations underlying the generally greater success of face-to-face interviews in generating more consistent data on sensitive behaviors than, for example, various types of self-administered questionnaires (Mensch et al., 2008a, 2008b).

These assumptions having been made clear, there are two recommendations related to questionnaires. First, it is useful to include questions that can allow for the estimation of different types of error. One, for example, might ask questions that allow analysts to identify respondents with a greater propensity to lie. Imagine a small series of questions of the form: "Have you ever heard of X?" where "X" is sometimes a well-known public figure (politician, sports player, singer) and at other times a completely made-up person. These are fast to administer and code (Yes/no answers).

Another small series might be directed at identifying the level of acquiescence bias, that is, the propensity of people to respond positively to any question. Some acquiescence bias might be partly captured in the "Have you heard of X?" questions. But it can also be captured more directly by reverse-ordering question wording on a small series of attitudinal questions in a portion of the questionnaires, then randomizing their allocation to the sample. (The cumulative difference in answers between the two groups is then treated as an indicator of the size of the acquiescence bias in the sample).

Finally, it is also desirable to include questions that allow for the estimation of social desirability bias. This is tricky since respondents' motives in answering questions are sometimes hard to pin down. But the best general approach is to embed such questions in a local event or program. For example, in an African setting in which there has recently been a national public health campaign promoting the building of pit latrines, the use of malarial nets, or HIV tests, an interviewer might ask the respondents whether they have had these things. Where the respondent answers "yes" the interviewer might find some way to verify (e.g., by asking to use the pit latrine sometime later in the interview; or by asking the respondent—or someone else in the household—whether they like sleeping under the net; or by collecting gross data on the number of HIV tests conducted at the local clinics).

The idea, in other words, is to find some way to validate responses to questions that deal with desirable behaviors, whether they are considered desirable by the authorities or by respondents' peers. And if that validation cannot be direct—as in directly observing whether there is a pit latrine or not—then at least it should be indirect. For example, if we can be reasonably sure that the direction of a response bias on a given question is downward (e.g., "have you ever had an abortion?"), then a particular data-collection mode which generates higher responses is considered to generate better data. Data-collection methodologists have used this type of indirect validation a lot, and very productively (e.g., Aquilino, 1994; Mensch, Hewett and Erulkar, 2003; Mensch et al., 2008a, 2008b; Plummer et al., 2004).

The second questionnaire-related recommendation deals with questionnaire translation, particularly common in many areas of the developing world (since the questionnaire is typically constructed in a global language, and many developing countries have multiple language groups). It is important to translate a questionnaire. Not doing so—even where interviewers have in the past used a translated questionnaire so are roughly familiar with the wording of the questions—increases interviewer's role-restricted error, which expresses itself in the real data as extra random error or noise around the mean (Weinreb and Sana, 2009b). The question is: how should one translate? Normal practice is to have one person translate, a second person back-translate, then somehow combine these two methods. The recommendation here is to have the translation conducted by a group who know what it is that you want to know. In the best-case-scenario, this group should be constituted by survey supervisors. Even after full training, translating the questionnaire will make them more deeply familiar with it. The more familiar they are with the questionnaire, the easier it will be for them to train interviewers and catch interviewers' errors in the field.

When the questionnaire is being translated however, questions over which the translators struggle or have debates should be flagged. For if it wasn't immediately clear to them how to translate what you meant, it is likely that the translation will not be completely clear to respondents either. And if the questionnaire has to be translated into several languages—as is often the case for large surveys in multiethnic/multilingual developing countries—but in only one or two of them are there heated debates about how to translate a central question, then this has implications for the reliability of group-level comparisons on that question. Remember, if the languages in which there were problems translating are known, then analysts can directly check

whether that is a source of the problematic comparison. If they are not known, those analysts are forced to guess.

iv. Interviewer selection

The selection, training, and oversight of interviewers are amongst the most important methodological tasks in the field. Interviewers are the eyes, ears, and mouths of the project. All data that we use is filtered through them, by the way they ask questions, by the way that they interact with respondents, and by the way that they code respondents' answers.

There are a number of general guidelines related to the selection of interviewers.

First, hire lots of interviewers. It is true that, administratively, it may be a hassle to manage them. But the financial implications are negligible—e.g., the number of interview days is the same whether one thousand interviews are conducted by fifty interviewers or by ten. Moreover, and this is the most important point, it makes much more sense methodologically to hire a lot of interviewers since doing so dilutes the "design effect" that can be associated with a bad interviewer—that is, the deviation from expected error associated with a simple random sample. For example, where a bad interviewer is responsible for 10 percent of the sample (ten interviewers conducting one hundred interviews each) his error is compounded across a much larger share of the data than where he is responsible for only 2 percent of the sample (50 interviewers conducting 20 interviews each). Hiring more interviewers is therefore the easiest way to avoid this (see Fowler and Mangione, 1990, for more on this effect). Finally, having a larger stable of interviewers means that it is easier to get rid of bad ones—reducing your interviewer pool from ten to eight has a lot more implications for the pace of fieldwork than reducing it from fifty to forty-eight.

Second, in most settings, worry much less about interviewers' social characteristics than you've been told to, or are in the habit of doing. For example, many projects habitually employ women interviewers, especially if there are health-related questions, or female respondents. While there is some empirical evidence to support this practice in Nepal (Stone and Campbell, 1984; Axinn, 1991), there is none elsewhere. In fact, all available analyses in Africa hint at the reverse, i.e., better information, or at least information that is as good, is given to male interviewers (Blanc and Croft, 1992; Becker, Feyisetan and Makinwa-Adebusoye, 1995; Weinreb, 2006). There appear to be no extant studies conducted in Latin America or elsewhere

in Asia. Consequently, the best policy in general is to hire the best-qualified interviewers whose first language is the language of the prospective respondents, irrespective of their gender, religiosity, class, and race. Having done that, do the following:

- Have each interviewer fill out a short questionnaire which collects basic sociodemographic data, including age, gender, ethnicity, marital status, religion.
- Randomly assign them to respondents—this issue is revisited below.

Third, where possible—in particular, where there is a limited number of field sites and a sufficient number of individuals with at least a high-school education around those field sites—use local interviewers. This has a number of advantages:

- In many LDC areas it will increase response rates and also generate higher-quality data (Weinreb, 2006).
- It embeds the project in the community, giving a stake in the project to every family in which one member is directly benefiting from employment. This, in turn, can help defuse any points of tension that may emerge during fieldwork.
- It saves the project money since there will be no need to house people.

Fourth, even if local interviewers are not used, make sure that you and your team select the interviewers. In many countries, interviewers receive their jobs because they are part of a permanent stable of interviewers maintained by the National Statistical Office, or through cronyism. Do not use these interviewers unless you have to (and if you have to, make sure that you can have them "reassigned elsewhere" if they are underperforming on your project). Instead, post ads in a local newspaper or on local notice boards that asks all people with a given characteristic who are interested in a temporary position as an interviewer/enumerator to come to an interview on a certain day (with their high-school graduation certificate). Reduce the size of this initial group with a written test. Have all test survivors interviewed to check their interactional abilities (people who score extremely high on the written test can be terrible interviewers, though in that case they are often very good data editors or data-entry clerks). Finally, remember the aim of these tests is not to generate a final pool of interviewers. Rather, it is to generate a starting group for interviewer training. Given expected attrition during training, make this group roughly 20 percent larger than your target team size.

v. Interviewer training

Specific training will depend on favored interviewing style. The recommendation here is that researchers consider "opening up" part of the questionnaire. That is, as implied by prior recommendations regarding questionnaire translation, a more conversational style of interviewing anchored by a structured (and translated) questionnaire is likely to generate higher-quality data (allowing analysts to draw more valid conclusions about the causes of change or stagnation in given indicators of interest). A more conversational style, however, also implies that interviewers need to be extremely familiar with the questionnaire, since only this way will they be able to probe or rephrase a misunderstood question in ways that are consistent with the question's intent.

The aim of interviewer training goes beyond this however. Aside from giving researchers and survey supervisors the opportunity to observe interviewer trainees in action (to see how effectively they communicate, how speedy and careful they are, how long they can remain focused, and how easy they are to work with), interviewer training also provides a final opportunity to improve the questionnaire. This is because interviewers, especially if they are local, are more likely to be familiar with the subtleties of local dialects and the latest in local norms, unlike survey supervisors, who are often university educated and live in urban areas. During training, interviewers can tell you where they think that the wording of a given question does not, in this local setting, communicate what it is intended to communicate. They may also feel good about you listening to their opinion, not merely treating them as employees.

All of these aims can be achieved by making training hard hitting. Set high goals, test trainees frequently, and eliminate those who are not doing so well (either technically or because they're too contentious and arrogant to be a good interviewer). This may include trainees with past experience working as an interviewer. In fact, they may be especially prone to failure if they were badly trained and/or are arrogant about their skills.

On the flipside, treat the trainees well. Provide little surprises (soft drinks, T-shirts at the end of training, etc.). Tell them that those who complete the project will receive a personal letter of recommendation from the project principal investigator that you hope will help them obtain other work in the future. For many, this will be a coveted reward (and, printed on headed notepaper from an international organization, a useful one).

vi. Interviewer assignments

The key principle here is to use an "interpenetrating" sampling technique that randomly assigns interviewers to respondents. That is, organize interviewers in teams of four to six and assign them, and only them, to a given sample cluster. Within that cluster, randomly assign those interviewers to respondents. Using an interpenetrating sampling technique is the only way to allow analysts to distinguish cluster-level differences from other sources of response variance.

For example, one of the most important sources of measurement error is associated with role-restricted interviewer effects. If only a single interviewer is assigned to a single cluster of houses, area, or village—common practice in many studies since it is administratively easier to organize—it will not be possible to distinguish area- and interviewer-specific sources of variance. In fact, not only does this data collection design make it impossible to identify the contribution of interviewer-related error; it also makes it very difficult to unambiguously identify whether differences between areas stem from real differences as opposed to differences among interviewers. Since the identification of area-specific differences is often a critical element in social research in general, and evaluation research in particular, this is highly problematic.

Randomized assignment of interviewers of given social characteristics (gender, age, etc.), is also the only way to reliably judge the extent to which those characteristics matter—that is, whether there are "role-independent" interviewer effects.

D. Conclusion

The data that are used in social research in general, or evaluation research in particular, emerge from particular social contexts. Within those contexts the data are also affected by a number of more traditional methodological factors—types of interviewers, questionnaires, etc. This chapter has attempted to lay out a relatively simple set of data-collection guidelines that seeks to minimize the extent to which those contextual factors and methodological factors affect data, causing it to deviate from its true value. Taking into account the varied interests—and incentives—of actors, both formally involved and merely interested, is the first step, since it provides researchers with important clues about constraints facing different players, and the social networks that they can influence. Likewise, acknowledging that levels of error are unknown and cannot be known a priori, and that post-hoc econometric methods are limited and problematic in their own right, prods us towards designing studies that allow for the independent

estimation of sources and scale of error in actual data. Given space constraints, an attempt has been made to cover the most important sources of error. In other words, the underlying assumption is that those elements that are not represented here—examples include selecting survey supervisors, setting up logging systems, organizing data entry and data cleaning—are less important sources of error. In addition, there is much less contentious debate among field workers about how to organize these elements.

One exception to this is the practice of gifting respondents. This is an emerging methodological concern in LDC data collection. Also, it is highly relevant to evaluations of conditional cash transfer programs and other programs where one group of respondents benefits from a program while a control group does not. The next chapter is therefore devoted to this topic.

Section 5. Incentives and their Effect

Evaluation research depends on good data. The previous section has detailed areas where substantial improvement could be made in order to enhance the quality of data used by researchers. The argument throughout, primarily based on empirical evidence from developed countries, is that there are important methods in existence to both evaluate and improve data quality. This section discusses one specific approach—gifting—that has been widely used in developed countries for improving data quality and which has been used somewhat haphazardly in developing countries as well. The section begins by explaining why gifting—or the provision of incentives to respondents—may offer an accessible and even affordable strategy for improving data quality. Subsequently, exploratory analyses of data from the PROGRESA poverty program are presented in order to provide some indication of how "indirect" incentives inherent in program participation may alter both survey participation and item nonresponse rates, as well as the quality of data that is obtained from those that do respond.

A. Theoretical explanations for why incentives might matter

At face value, paying respondents for their participation in a survey seems like a very reasonable practice. Such payments have been used for a long time and there are strong theoretical reasons why gifting should improve the quality of responses obtained in the course of surveys. Contrary arguments can also be made, however, showing how providing gifts to respondents might reduce data quality. It is useful to review these different arguments.

In traditional economic terms, it is easy to see gifting as a type of narrowly focused economic exchange geared toward increasing an individual respondent's motivation to participate in a survey (Datta, Horrigan and Walker, 2001). That is, the incentive could increase a wavering respondent's motivation to participate in the study. It could also buy more honesty—pushing respondents to provide better, more accurate answers, or divulge more personal information than they otherwise would have.

An alternative basis for arriving at the same prediction builds on well-known theories of social interaction. A general theoretical perspective would treat gifting as an essential part of everyday social interaction that, along with exchange in general, embeds people in what Joseph (1993: 119) calls "webs of relationality" (see also Schwartz, 1967; Green and Alden, 1988).

When relations between two people are already established, that is when they are familiar to each other, gifting adds another layer, cementing the relationship a little more.

However, for people who do not know each other—"strangers" in Simmel's (1950: 404) classic essay, and a category that includes survey interviewers – gifting can be seen as more than an attempt to motivate the respondent to consent to be interviewed (the economic rationale). It is also an attempt to establish a relationship in which the stranger comes closer to the status of "insider,"— someone known and trusted —at least for the duration of the interview. Achieving a sort of insider status, it is hoped, will increase the likelihood that a potential respondent will participate and then, once they have agreed to be interviewed, will do their best to provide accurate responses. It follows from this that both participation rates as well as the reliability and validity of responses may differ depending on whether the respondent has been paid or not.

The contrarian perspective provides what, on a priori grounds, appear to be equally compelling explanations for why gifting may be better avoided. One argument is related to the above-mentioned notion of gifting and "webs of relationality." If gifting entangles a respondent more closely with the interviewer, it may also raise the likelihood that the respondent will attempt to offer biased answers aimed to please the interviewer.

The motivation to respond in this way—that is, with a social desirability bias—intensifies when data are collected through face-to-face interviews, one reason that self-administered questionnaires may be preferable on sensitive topics (Tourangeau and Smith, 1996). For example, Presser and Stinson (1998) reexamine data on trends in religious attendance in the United States over several decades. Whereas earlier studies using interviewer-based surveys showed little decline, analysis based on self-reported questionnaires showed a strong trend towards secularization. The authors claim that respondents preferred not to divulge their real behavior to interviewers, leading to a distorted trend in the data (Presser and Stinson, 1998). Another line of research has focused on how the joint characteristics of the interviewer and the respondent may lead respondents to bias their answers to avoid displeasing or to please interviewers (Groves, Cialdini and Cooper, 1992).

Another argument against gifting is more closely economic in nature. This perspective, which initially built on experiments in social psychology, suggests that an incentive may unintentionally reduce the effort invested in an activity. Titmuss (1970) demonstrated a version

of this concept using data on blood donations that showed that payment for blood led to actual declines in donations.

To economists, the explanation is that payment reduces the gains that individuals receive from altruistic behavior (Frey and Oberholzer-Gee, 1997). In one example, the authors showed how experimental participants were less likely to respond positively to a request to house a nuclear facility in their own community if they were offered compensation. This general argument, that individuals are less likely to fulfill their civic duty when paid, is the subject of considerable testing.

Gneezy and Rustichini (2000) found that subjects answered fewer questions correctly on an IQ test if they were paid a small fee per correct answer. The social-psychological explanation is that an individual's intrinsic motivation for performing tasks will decline when they perceive a controlling external intervention that is controlling (Deci 1971). The literature in social psychology has demonstrated that extrinsic rewards could reduce individual motivation for certain activities (Deci 1971; Kruglanski, Friedman et al. 1971). Mapping this general result onto the question of survey participation or quality of responses generates a simple expectation: one would expect that an individual's motivation to respond to a survey would decline when offered an incentive. This could lead both to lower participation rates and lower quality data for actual respondents.

There are, of course, ways to complicate each of these accounts, those that are pro-gifting and those that are anti. For example, it may be that the effect of gifting may depend on the type of gift (money, a service, a bag of sugar, a book token, etc.), when it is given (at the beginning of the interview, at the end), how the gift if presented (as payment, as a "token of appreciation") on where the giving takes place (wealthy suburb, poor shanty town, isolated rural area), and so on. Not surprisingly, neither existing theory nor existing empirical record provides a clear way to predict how gifting will impact data quality from surveys across all such contingencies. Consequently, the following subsection looks at empirical evidence to help untangle this relationship.

B. Existing evidence on the impact of incentives on survey response

Gifting has long been one of the standard tools used by researchers to protect the external validity of sample surveys (by reducing selectivity bias associated with survey nonresponse).

This is proving particularly important over time as resistance to survey response appears to be increasing in the United States and other developed countries (Singer et al. 1998; Groves and Couper 1996). In MDCs it has been shown to increase the willingness of respondents across different types ("modes") of data collection, i.e. face-to-face, mail, phone, medical examinations, expenditure diaries (Ferber and Sudman, 1974; Willimack et al., 1995). Studies on this relationship go back to the 1970s and before. Ferber and Sudman (1974) report on a number of experiments intended to gauge the effects of incentives that were conducted in the 1960s and 1970s and which showed generally positive effects of incentives on participation rates.

More recently, a study by Willimack et al. (1995) reports on an experiment conducted where incentives were given to a random sample of respondents in a face-to-face survey. In this case, a non-monetary incentive (a ballpoint pen) raised response rates by about 5 percent over the control group. Church (1993) provides evidence from a meta-analytic study of some 38 experimental or quasi-experimental studies on the effect of incentives on mail survey response rates. The findings indicate that incentives that provided immediate rewards to respondents raised response rates. The findings also indicate that promised rewards upon completion offered no benefits for response rates. This latter point appears to reflects a consistent finding—incentives must be paid out immediately to be effective (Berk 1987).

An area of growing interest is the effect of incentives on panel survey respondents. It is particularly important to reduce attrition on panel data, which are expensive to begin with, and incentives have been suggested as one potential tool. Panel studies have become much more common in the US and other developed countries. Panel surveys are also increasingly in use in developing countries. This is particularly the case for evaluation research where the panel-experimental survey design has become the gold-standard in evaluation.

Overall, the evidence on gifting and attrition from developed country surveys appears relatively consistent: it indicates that panel attrition may be reduced if incentives are used (Ferber and Sudman 1974). More recently, Zagorsky and Roton (2008) show that an experiment by the US government to increase participation by offering incentives to those that previously refused was very successful in raising response rates in the National Longitudinal Survey of Young and Mature Women. Interestingly, they did not find any clear effect of incentives on participation rates for those that had responded in the prior round. A recent experiment based on the British

Household Panel Survey showed that raising the incentive from 7 to 10 pounds per interview led to a substantial increase in response rates.

Incentives have not only been shown to affect survey response rates but also appear to have an impact on data that is obtained conditional on participation in the survey. While there is concern that incentives may reduce data quality, studies have not generally shown this to be the case. The early review by Ferber and Sudman (1974) of panel expenditure surveys indicated that compensation in some form was associated with more complete and accurate responses. In the Willimack et al. (1995) study mentioned above, measurement error did not appear to increase or decline as response rates rose.

Singer et al. (1998) provide evidence that incentives had no negative impact on data quality. In this case, their assessment of response quality is primarily based on the frequency of nonresponse or "don't know" responses to a series of questions. These nonresponses are taken as an expression of intensity of effort by the respondents. Also, the respondents that received incentives also appeared to express more favorable attitudes towards similar surveys. In another study, Singer et al. (2000) use data from the monthly phone Survey of Consumer Attitudes while they find no clear effect of incentives on participation rates, they do note a significant improvement in data quality with item nonresponse rates lower for those receiving incentives.

A cautious interpretation of these findings would suggest that when incentives are actually provided in person or in advance then they appear to raise overall response rates. Furthermore, there is no indication that data quality declined—in fact, where there is an effect, it appears to be positive. However, despite this promising record of empirical evidence, which is primarily based on experimental data, the ability to infer the implications for data collection strategies in developing countries remains limited at best.

In fact, in relation to LDCs, only non-experimental evidence can be brought to bear on this question. Most of this is anthropological and addresses lone researchers who are attempting to establish themselves in communities through participation in local exchange networks (e.g., Agar 1980, Barley 1983, Bernard 2000). In relation to survey research, some of the key themes that emerge in these texts have already been covered in fieldwork recommendations (Section 4). In particular, the recommendation to give locals a stake in the project—for example, by employing them in some capacity on the survey, or by purchasing supplies locally—are motivated in part by the gifting/exchange principle. The same is true of the recommendation to

give the respondent a copy of their photo. Somewhat more intrinsically connected to survey response: the only extant paper that explicitly addresses this issue in an LDC setting compares survey attrition between a first and second round of data collection in two comparable research projects (Weinreb, Madhavan and Stern 1998). The first, in a rural area in Kenya, did not gift; the second, in a rural area in Malawi, did. Not surprisingly, attrition was lower in the Malawian sample, but again, the ability to extrapolate from this single result is severely limited by the non-experimental nature of the study.

C. Making Use of the PROGRESA Conditional Cash Transfers Experiment

The ideal experiment to test the impact of gifting would be one where the incentives are randomly assigned to recipient households. The random assignment would ensure that any change in the survey response rate is due to the provision of incentives. Such experiments, cited in the previous section, have been carried out in the US and a few other Western nations, but none appear to have been carried out in LDCs. the following analysis therefore makes use of an alternative research design—one where households are randomly chosen to enter a program where they do receive incentives. These are conditional incentives, but the large majority of eligible households in the treatment group receive them. A major advantage of this design is that it uses data from PROGRESA, a huge evaluation project with an enormous sample size. Furthermore, while the randomization is at the community level, within-community differences in gifting allow for the exploration of neighborhood deprivation effects—that is, how individuals' own response behavior is affected by their neighbors' receipt of incentives.

i. The PROGRESA Data

PROGRESA is the largest and best known conditional cash transfer programs instituted in Latin America in the mid- to late-1990s. PROGRESA began in Mexico in 1997 as a mechanism for addressing extreme rural poverty. The PROGRESA program focuses on the development of human capital of poor households by improving education, health and nutrition outcomes. Because PROGRESA targets poor households, criteria were developed for determining eligibility based on household well-being and selection (Skoufias, Davis and De La Vega 2001). This process involved three separate stages aimed at identifying potential recipient communities, determining eligible households from within those communities, and finally to having the

selections reviewed by local experts. Because the eligibility criteria are based on poverty, eligible households are referred to as "poor" in subsequent sections. The random assignment was based on a census in 1997. Communities were then randomly assigned to either treatment or control groups with treatment beginning in early 1998. Skoufias (2005) provides a very thorough overview of the PROGRESA program.

Two different forms of cash transfers are provided to households to meet these objectives: a food grant and a school scholarship. Each component is linked to separate and independent conditionality requirements described below. In both cases and with rare exception, transfers are provided directly to mothers under the assumption that they are more likely to use the resources to benefit their family and children. The first incentive is a food grant, which is the same amount for each beneficiary household (US$16 per month as of 2001), and is conditional on health check-ups for all family members and attendance at public health lectures. The second set of incentives is tied to schooling. School scholarships are linked to specific children and thus differ by household. The grants are awarded to mothers every two months during the school calendar year and all children over 7 and under 18 (during this period for grades 3 through 9) are eligible. Children must register and ensure regular attendance (a monthly attendance rate of 85%) to receive the award (Adato et al. 2000).

By the end of 1999, the year corresponding to the data in the sample that is used, PROGRESA provided bimonthly transfers to approximately 2.3 million households or about 40 percent of all rural families and 11 percent of all Mexican families. With the advent of the Fox Administration in 2001, PROGRESA changed its name to OPORTUNIDADES and expanded operations to urban and semi-urban areas—until then it had been limited to communities with less than 2500 inhabitants. The PROGRESA budget for 2002 reached US$1.9 billion, covering almost three million rural families and over 1.2 million urban and semi-urban families (Fox 2002; Skoufias and McClafferty 2001).

The present analysis uses data from the March 1998 and the November 1999 datasets. The March 1998 data were collected right at the onset of the program before households would have received any benefits from the program. The 1999 data reflect a later period where most households had begun to receive the benefits of the PROGRESA program but control households had not yet been incorporated into the expanded Oportunidades program. Once the Oportunidades program began, the clean experimental design initiated in PROGRESA ended.

Studies based on Oportunidades rely on matching designs but these two rounds offer the best opportunity to make use of the randomized, experimental design.

The analysis distinguishes between four groups naturally created by the experimental design and the eligibility criteria that are part of the program. These four groups are the product of the project's two-stage sampling design. First, communities were assigned at random to treatment (receive incentives) and control (do not receive incentives) group. Second, within the treatment group, poverty criteria were used to establish eligibility for given households. The four groups are therefore:

1. Treatment group, poor (receive transfers)
2. Treatment group, not poor
3. Control group, poor (does not receive transfers)
4. Control group, not poor

Finally, note that information was not available to assess participation rates at the first census. This critical failure means that selectivity bias created by nonparticipation at the earliest stages of the project cannot be investigated. In turn, this means that the analysis is essentially limited to nonparticipation in later waves contingent on first participation. This can be interpreted as a lower bar to gifting effects.

ii. Empirical Analysis of Program Incentive Effects on Responses

Initial focus is on three sets of preliminary results obtained by analyzing the survey data from PROGRESA. The first focuses on participation behavior by respondents. The second focuses on item nonresponse, which is conditional on individual participation in the survey. The last part utilizes actual responses obtained on relatively sensitive questions where responses might reasonably be influenced by incentives.

Estimates are based on a difference-in-difference (DD) approach where data are available from both the 1998 and 1999 rounds of the survey. The DD approach is a simple and intuitive statistical design for analyzing panel data with a randomized treatment and control groups. Other included variables are a dummy variable for treatment to capture whether a household is in the treatment (treatment=1) or control (treatment=0) group, a dummy variable for year to capture whether the measurement is from March 1998 (year=0) or November 1999 (year=1), and an interaction between treatment and year. Thus, the model takes the form,

$$Outcome = b0 + b1*(year) + b2*(treatment) + b3*(year*treatment)$$

The treatment and year interaction, captured with b3, provides the DD estimator providing a convenient coefficient for the amount of change in the outcome over time for the treatment group minus the change in the outcome for the control. The statistical test on the interaction coefficient, b3, provides a simple statistical test for whether this coefficient differs from zero.

Alongside the coefficient estimates, the corresponding transition in probabilities associated with each probit estimate are also presented. Because the main explanatory variables are all dummy variables, the results presented are those of the corresponding changes in probability when the dummy shifts from a value of "0" to "1". Note that additional controls are included for head's age (in three categories), years of education (in three categories), sex and whether the head is indigenous. These controls are included in all models but are omitted from the tables. The estimated probability values are calculated at the mean of all other variables.

Two other points should be noted. First, use of a nonlinear estimator such as the probit with interactions terms may lead to incorrect coefficient estimates and significant tests (Ai and Norton, 2003). However, this problem is not acute in these data, since the groups do not fall in very different regions of the probability density function. Nevertheless, as a test of robustness, main models were rerun in a linear probability model. This model generated substantively similar findings, signaling their overall robustness.

Second, in several instances the analysis is limited to a single round of data. Where this is the case, it means that analysis is based on cross-sectional differences between treatment and control communities. While the panel is preferable because differences may always exist between the treatment and control prior to treatment, the cross-section nonetheless makes use of the randomization inherent in the random assignment to groups. Furthermore, in both the panel and cross-section, eligible (poor) respondents are distinguished from ineligible (nonpoor) counterparts.

These analyses are presented for demonstrative purposes and are not intended to offer the most accurate estimates possible. For example, the standard errors are not corrected for clustering produced by the sample design, a procedure which is highly recommended in actual empirical work and which will likely increase the standard error estimates shown. Thus, the

coefficient estimates can be seen as accurate but their significance levels should be viewed cautiously.

a. A basic participation model

Results are presented in Table 5.1. The first set of estimates includes all households. The second and third were run separately on eligible (poor) and ineligible households. Overall, the results show that households in the treatment communities at the onset of the program had lower response rates but that this difference is not significant (p=0.11). The coefficient on year indicates a very significant decline of about 2 percent in the participation rates for the control group between the two rounds. It can also be seen that the decline over time in completion rates for the treatment group is more than 3 percentage points larger relative to the control group, or nearly twice as large. This coefficient, which is highly significant, underscores the very large decline in participation that occurs in the treatment communities.

Table 5.1 Probit Estimates of Survey Participation for All Households and by Eligibility (Poverty level), Difference-in-Difference using both 3/1998 and 11/1999 PROGRESA Data[1]

	All Households		Eligible Households		Non-Eligible Households	
	Completed Interview	Prob. Δ	Completed Interview	Prob. Δ	Completed Interview	Prob. Δ
Treatment	-0,038	-0,007	-0,006	-0,001	-0.079**	-0,014
	0,024		0,032		0,035	
Year	-0.098***	-0,018	-0,021	-0,004	-0.186***	-0,034
	0,026		0,035		0,038	
Treatment x Year	-0.166***	-0,032	-0.145***	-0,028	-0.184***	-0,035
	0,032		0,044		0,047	
Constant	1.015***		0.913***		1.113***	
	0,03		0,039		0,047	
No. of cases	48034		24998		23036	
p<0.10, ** p<0.05, *** p<0.01						

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

A more informative perspective on survey participation is gained by looking at the association between participation and eligibility—since eligibility is also a proxy for wealth.

Although only poor households are eligible for the cash transfers, wealthier (ineligible) households are also included in each round of the survey. When the focus is exclusively on households that were eligible for treatment, no difference between treatment and control groups can be observed at the onset. Nor is there evidence of a significant decline over time in participation for poor, eligible households.

However, the interaction term, the most interesting coefficient in this model, indicates that the decline in response rates for the eligible households in the treatment group is 2.8 percent greater than the decline for eligible households in the control group. This coefficient is consistent with expectations that participation and benefits reduce the incentive to respond—at least relative to households that may be hopeful (rightly so it turned out) to be incorporated into the program at a later date. It seems that households in the treatment group do not see the incentive as at risk in comparison to households in the control who may feel that nonparticipation may endanger their future receipt of benefits.

The results for ineligible households are also of interest. Here, it is noteworthy that even though ineligible households exhibited lower participation rates already in 1998, there is a decline in participation for control households of some 3.4 percent between the two rounds, and that the decline for ineligible treatment households is larger by another 3.5 percent. Thus, the program incentives appear to exhibit spillover effects with particularly strong declines for households that are ineligible but in the treatment community.

The decline in their participation may not be terribly surprising given that they have continued to be subjected to the survey but also have seen many of their neighbors benefit from the program. Again, to judge the robustness of this result, these two models were also jointly estimated in a triple-difference model (that provides a test statistic to compare the poor and nonpoor differences in the change in participation over time). This result is not shown here, but it is noteworthy that the triple-difference estimator is not actually significant. This is not surprising given that both are strongly negative. Thus, while the effect of treatment on participation is larger for the poor in 1999, the difference in its effect across the two years is not statistically different ($p=0.54$).

Further analyses of the reported causes of nonparticipation, which are not reported here, show that absolute refusals increased in the treatment group relative to the control group. Interestingly, the probability of a claim that the household relocated also grew in the treatment

group more than in the control group. The claim of relocation is not very plausible, both because fewer relocations can be expected given the incentives inherent in staying in the treatment communities, and because there is some evidence showing declining out-migration in treatment households (Stecklov et al., 2005).

b. Item nonresponse

The possibility that households selectively answer questions raises serious concerns. On the one hand this has long been recognized, as indicated by the efforts to develop reliable imputation methods for use with item nonresponse. On the other hand in much applied research, the selective answering of questions is often overlooked. In particular, while participation decisions are frequently examined, the selectivity introduced by item nonresponse can be equally problematic for many interpretations. In order to highlight some of those problems, the following analysis focuses on reported current pregnancy status in the March 1998 round of the data.

At initial look at the data already raises concern. Out of the 19,148 women that were supposed to have been asked about their current pregnancy status, there is missing information for 1,250 women. Because this question was only asked in a single round, analysis uses a simple model with poor, treatment and their interaction.

The results—presented in Table 5.2—highlight how serious some potential biases may be, whether or not they are affected by the incentives introduced by PROGRESA. For while on the one hand the likelihood of providing no answer or saying "don't know" is similar for treatment and control households—that is, there is no evidence that the program incentive alters the effort or desire to offer a response—on the other hand, there is a very large and negative effect of being poor. This is surprising because the nonpoor should know their status at least as well as the poor.

Two explanations can be used to explain this coefficient. One is that nonpoor households are less prone to sharing their pregnancy status with interviewers. The other is that interviewers may feel more comfortable "pushing" respondents in poorer households towards providing a response—regardless of whether a response is known by the respondent or voluntarily given. (Assuming that interviewers are assigned using interpenetrating sampling techniques, this type of interviewer effect could be assessed by using approaches discussed in Section four). In either case, there is clear evidence of a response effect varying by socioeconomic status (SES) which,

as indicated in Section three, is contrary to the standard assumptions on which econometric fixes depend.

Table 5.2: Probit Estimates of No Response to Question on Current Pregnancy Status, March 1998 PROGRESA Data[1]

	No Response	Prob. Δ
Treatment	-0,044	-0,003
	0,04	
Poor	-0.276***	-0,018
	0,055	
Treatment x Poor	-0,018	-0,001
	0,069	
Constant	-2.081***	
	0,085	
No. of cases	20398	
p<0.10, ** p<0.05, *** p<0.01		

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

c. The quality of responses to potentially sensitive questions

The last set of analyses focus on examining the quality of responses actually provided and whether the program incentives may be an influence. In other words, here analysis shifts from the participation decision to whether and how people answer questions. Of particular interest are questions that touch on potentially sensitive behaviors—though any answer is of course conditional on two prior decisions: having agreed to participate in the survey and having provided any type of answer.

The first of these sensitive questions focuses on whether or not individuals report having had an abortion or lost a pregnancy over the past period (Table 5.3). This question was only asked of nonpregnant women. A series of subsequent models (Tables 5.4A, 5.4B, 5.5A and 5.5B) examines recent cigarette and alcohol expenditures. Analysis follows along the same lines as above—focusing both on the treatment process as well as on the effect of eligibility (wealth). It begins with a set of models demonstrating the impact of treatment, year and their interaction (treatment x year) on whether an individual reports any cigarette expenditures for the prior week (see Table 5.4A) and then a similar model is subsequently presented to show how these same variables influence the amount spent on cigarettes in the past week (see Table 5.4B).

Subsequently, Tables 5.5A and 5.5B show similar specifications but where responses to alcohol expenditure questions are tested.

Table 5.3 Probit Estimate of Reported Abortion or Pregnancy Lost, PROGRESA March 1998 Data[1]

	All Households	Prob. Δ
Treatment	-0.129*	-0,003
	0,068	
Poor	-0.279***	-0,018
	0,084	
Treatment x Poor	0.231**	-0,001
	0,104	
Constant	-2.095***	
	0,107	
No. of cases	17272	
p<0.10, ** p<0.05, *** p<0.01		

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

The test for the effect of treatment and eligibility on reported abortion (or lost pregnancy) certainly falls in the class of sensitive questions. Unfortunately, the question is only asked in a single round—March 1998. However, several interesting results emerge from the analysis. First, households in the treatment group that are nonpoor are less likely to report an abortion. This is important since it is consistent with the argument that respondents in ineligible households are less likely to share intimate or personal information with interviewers—the absence of a cash transfer means that no connection has been forged with them. Results also show that poor households are less likely to report abortions. Finally, the interaction term treatment x poor, which is significant, indicates that the effect of being poor (or eligible and actually or likely to be getting benefits) leads to higher reports of abortions or lost pregnancies. Again, this result supports the idea that providing incentives increases respondents' willingness to report on sensitive behaviors.

The next set of analyses focus on expenditure-related questions. The model on cigarette expenditure, estimated with a probit, shows a decline in the probability of smoking expenditure over time (see Table 5.4A). Interestingly, for smokers, results show that the amount spent on

cigarettes in the past week rises between the two rounds (Table 5.4B), although we find no difference in cigarette expenditures between treatment and control groups in either round.

Table 5.4A Difference in Difference Estimates of Any Cigarette Expenditure in Past Week[1]

	Any Smoking Last Week					
	All	**Prob. Δ**	**Poor**	**Prob. Δ**	**Non-Poor**	**Prob. Δ**
Treatment	-0,018	-0,002	0,003	0	-0,037	-0,003
	0,027		0,038		0,038	
Year	-0.262***	-0,025	-0.268***	-0,026	-0.257***	-0,025
	0,033		0,047		0,046	
Treatment x Year	-0,066	-0,006	-0,084	-0,008	-0,045	-0,005
	0,043		0,061		0,061	
Constant	-1.508***		-1.527***		-1.490***	
	0,021		0,03		0,029	
No. of cases	42957		22285		20672	
p<0.10, ** p<0.05, *** p<0.01						

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

Table 5.4B: Difference in Difference Estimates of Past Week Cigarette Expenditure Conditional on Spending[1]

	Cigarette Expenditures Last Week		
	All	**Poor**	**Non-Poor**
Treatment	-0,87	-0,144	-1,561
	1,141	1,799	1,411
Year	6.908**	8.851***	5.181***
	1,475	2,364	1,791
Treatment x Year	-0,473	-4,712	3,894
	1,944	3,078	2,391
Constant	11.057***	10.621***	11.377***
	1,75	2,524	2,568
R-sq	0,029	0,016	0,046
No. of cases	2132	1075	1057
p<0.10, ** p<0.05, *** p<0.01			

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

The model for alcohol expenditure highlights some intriguing findings (Tables 5.5A and 5.5B). A decline in the probability of reporting any alcohol expenditure from the past week can be seen for all groups. This effect is strong and consistent. However, even more interesting is

that the decline is stronger for the poor group. For the nonpoor, there is no interaction between treatment and year. Thus, those in the treatment group appear to report a strong decline in any alcohol expenditure over time. The coefficient here is similar in direction to the one on cigarettes, but the coefficient is highly significant in Table 5.4.

Table 5.5A: Difference in Difference Estimates of Any Alcohol Expenditure in Past Week[1]

	Any Smoking Last Week					
	All	**Prob. Δ**	**Poor**	**Prob. Δ**	**Non-Poor**	**Prob. Δ**
Treatment	0,008	0,0006	0,039	0,0032	-0,028	-0,002
	0,026		0,036		0,039	
Year	-0.631***	-0,049	-0.595***	-0,05	-0.676***	-0,048
	0,039		0,052		0,059	
Treatment x Year	-0.117**	-0,0085	-0.196***	-0,015	-0,019	-0,0013
	0,051		0,069		0,077	
Constant	-1.660***		-1.733***		-1.588***	
	0,043		0,056		0,071	
No. of cases	42853		22252		20601	
p<0.10, ** p<0.05, *** p<0.01						

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

Table 5.5B: Difference in Difference Estimates of Past Week Alcohol Expenditure Conditional on Spending[1]

	Cigarette Expenditures Last Week		
	All	**Poor**	**Non-Poor**
Treatment	-4.128**	-3,592	-4.309*
	1,666	2,184	2,597
Year	-13.589***	-13.365***	-12.329***
	2,922	3,743	4,728
Treatment x Year	10.926***	12.559**	7,034
	3,911	5,056	6,219
Constant	31.995***	31.460***	30.188***
	2,901	3,634	5,014
R-sq	0,012	0,006	0,016
No. of cases	-4.128**	-3,592	-4.309*
p<0.10, ** p<0.05, *** p<0.01			

1 Control variables not shown but include Head's Age in three categories, Head's education in three categories, Indigenous household, and Head female.

Alongside the decline in the probability of alcohol expenditure, there is also a notable decline in the amount spent over time for all groups. However, the positive coefficient on

treatment x year indicates a more muted decline in expenditure for poor households in the treatment group relative to poor households in the control group. In fact, the treatment group almost cancels out the time effect. This effect is not similar for the nonpoor, where the effect is in a similar direction but insignificant. Again, there are no effects for the nonpoor. This effect then is in the opposite direction of the one found in Table 5.4, related to abortions. Whether this difference is related to the fact that the abortion question reflects issues that are associated with the program as opposed to the alcohol consumption is not clear.

D. Concluding Remarks on the PROGESA Analysis

PROGRESA was not designed to experimentally test the effects of gifting on survey participation or response. The results presented here, therefore, cannot be seamlessly interpreted as evidence of the role of gifting on responses from developing country surveys. However, given the almost complete lack of any other evidence, the advantages of drawing on a very large and randomized evaluation, and the fact that households that are eligible (poor) and in the treatment communities could enjoy considerable benefits—equivalent to 20 percent of expenditure (Skoufias, 2005)—the results here are both instructive and important.

First, the results are suggestive of several relatively understudied aspects of gifting. One is that the benefits of incentives—if they exist—may quickly dry up. Once guaranteed a benefit, there is little evidence that respondents remember the gift. In fact, eligible households appear to show a much stronger decline in response rates over time if they are in the treatment group. Taken at face value, this result suggests that if gifts are given they should be given at the same time as the interview and that future rounds of interviews in panel surveys should gift again. However, it is worth noting that this result's significance is sensitive to clustering.

A second, perhaps more intriguing result, addresses the issue of unequal gifting. Even if rooted in people's differential wealth, the results presented here are consistent with arguments that unequal gifting may be seen as unfair, triggering an instinctive "I'm no sucker" reflex that makes individuals less likely to participate. This is the most likely explanation for the fact that the ineligible in the treatment communities experienced a particularly steep decline in participation.

Additional results are also worth noting. They illustrate on the one hand that there are strong differences in people's willingness or ability to answer questions on sensitive topics (e.g.,

pregnancy) and that these differences are tied to socioeconomic levels in Mexico rather than to program incentives. Furthermore, actual data responses also appear to be sensitive to treatment effects. In particular, it may be that individuals in treatment groups were more willing to respond to questions about abortion because they were gifted. In this case, the program incentive appears associated with muted declines in alcohol expenditures, although this could also be an income effect. Overall, for both treatment and control groups and for both cigarette and alcohol expenditure, there are overall declines in the proportion of people spending, but an increase in the amount that is spent by those who remain consumers. This may simply signal rising prices of such goods. There is no evidence in this case either that households in the program are strategically responding to questions appearing to spend less on alcohol. In that sense, the program results are surprising.

A more important implication of these findings is to raise a note of caution for researchers making use of data—however well known those data are. Gifting and incentives can have complex impacts on survey data that are reflected in participation rates, item nonresponse rates for specific questions, and the direction of responses for questions that are asked. These incentives or program effects then have the potential for creating unanticipated data problems in evaluation studies. If the program itself creates selectivity or measurement problems by generating more or less participation among respondents, or better or worse response quality, then the pure experimental treatment effect may no longer offer a perfectly valid approach to analyzing the data; or at least its validity, like the validity of data generated in other types of research design, is conditional on the level of survey error.

Finally, the results here are also intriguing in that they suggest that programs may themselves create nonresponse biases. In other words, program effects and response biases may themselves be intertwined, offering yet another threat to the interpretation of program effects. The extent of the bias will depend on the scope of the incentive effects relative to program effects. Panel data and difference-in-difference models do not help this situation. And while differencing can help to eliminate prior differences between treatment and control groups, the program incentive effects are caused by the program itself, making them indistinguishable. Note that these results, in particular the program incentive effects on the ineligible, differ from those identified in a recent paper on spillover effects (Angelucci and De Giorgi, 2009). While spillovers offer an interesting and original use of treatment effects, the use of the ineligible may

exacerbate the same types of data problems dealt with here. In particular, the results suggest that the ineligible in the treatment community are particularly prone to systematic biases or nonresponse. Interpreting a 10 percent increase in the consumption of ineligible households as arising from treatment could in theory confound—at least in part—spillover effects in data quality with spillover effects of consumption. Clearly, more work is warranted on these effects—in PROGRESA as well as other evaluation survey research. But this in itself is one of the recommendations to emerge from this report, a point to which the next and final section of the report is devoted.

Section 6. Conclusion

Impact evaluation is recognized as one of the most important stages in ensuring high quality and effective programs. Collecting data for evaluation is a critical first step in a process and this process must be incorporated into all program design. Yet, once the decision to collect data is taken, efforts must be made to ensure that evaluation based on the data can provide effective and reliable indicators for improving program performance. What can be done to improve the quality of survey data used for evaluation?

In this report, some key problems are highlighted that stand in the way of collecting good quality empirical evidence for evaluation as well as other forms of survey-based empirical research. They extend far beyond the two issues that evaluation researchers have traditionally focused on: how to sample and how to deal with data problems ex post, using econometric methods. This is, it can be argued, far too limited a focus. It ignores the multiple sources and layers of error that can accumulate in a given dataset, the lack of a priori knowledge about these levels of error, and the problematic assumptions that confound most attempts to address them. Moreover, in ignoring these things, it imperils the goals of accurate impact evaluation.

Also, while these problems may affect approaches to evaluation in general, they are particularly important in impact-evaluation research on developing countries, since it is precisely here that the sources of error are so little understood, and that the relevant literature is so sparse. Yet herein lies the true problem. It is in these areas of the world that evaluation research is so important, and increasing attention is now paid to them. Both general surveys such as the LSMS and the DHS now provide a wealth of information for researchers to use when monitoring conditions and studying development problems across a wide array of developing countries. Indeed, Grosh and Glewwe (1996) report that 30 percent of publications in the three top development economics journals were based on household survey data in 1995, a five-fold increase over the situation in 1975. Moreover, evaluation research is also increasing, not only in number, but also in sophistication, with random assignment, multiple waves of data collection, and moving beyond survey data alone into biomarkers, geocoded data, and similar.

By making an example of the PROGRESA evaluation this report shows that even high profile, gold-standard evaluation projects are not immune to the types of data-quality problems and effects discussed here. While the discussion presented here suggests mechanisms by which

incentives may lead to both better and worse data, the findings raise a number of red flags for evaluation researchers. These red flags are just as critical for those undertaking or using gold-standard type randomized evaluation data as they are for those using nonrandomized evaluation data. Some of the red flags might be made less problematic if more effort were made to reduce response effects in data collection. Section four of this report highlights a number of important guidelines that may help to improve data quality at all levels of the data-collection process. Only some of these, as pointed out, are explicitly addressed in the methodological literature. But all have effects on measurement error so need to be addressed in carefully designed studies.

Figure 6.1 Top Ten List for Helping to Reduce Survey Measurement Error at the Source

1. Follow basic administrative guidelines (4A)
2. Clarify the "central players" in the region and nationally and be certain to consider ways to work with them and reduce the chance of "spoilers" (4B)
3. Conduct pre-testing of all questionnaires (4C)
4. Hire relatively large numbers of interviewers, whom should be tested in the course of training, while setting high goals and providing rewards for success. (4C)
5. Interviewers should be assigned using interpenetrating sampling techniques (4C)
6. Consider all potential errors of non-observation including sampling, coverage and nonresponse (4C)
7. Include questions that can allow ex-post identification of different types of errors of measurement (4C)
8. Carefully evaluate whether there are systematic nonresponse patterns that might affect interpretation of findings (5)
9. Design clear guidelines for filling in missing data, preferably using interviewer teams done shortly after each day of data collection
10. Attempt to compare results to those that might be obtainable from routine statistics or other different data sources

Ultimately—and perhaps this is the central point—evaluation data itself needs to be evaluated. The best approach is to seek validation sources for the data collected in the surveys. The use of validation studies has greatly expanded our confidence in data-collection systems in developed countries. However, validation efforts are sorely needed for developing countries. In particular, large-scale impact-evaluation surveys might already incorporate such mechanisms

(and possibly resources that are necessary) to collect data to be used for validation. This is certainly not possible in all developing countries, given that valid data may simply not be available, but in some countries this is certainly an opportunity that should be sought.

A possible interim strategy is to focus a great deal more attention on the pre-test. Rather than use the pre-test as an informal test of the survey instrument, it should be used as a more substantial test and refinement of the entire data collection process. That means beginning with the questionnaire design, through to respondent participation decisions, interviews and social interaction in the interview process, and proceeding forward through data processing and editing. Measurement error can be introduced at each stage and an expanded pretest process, where respondents are sampled rather than chosen through convenience, may offer a tool to refine the data-collection process. Even if the pretest subsample cannot be incorporated in the final sample, then the value of this process may outweigh its cost. A smaller sample size may mean slightly larger sampling error. However, the total survey error may be reduced by this procedure, even if it remains more difficult to identify.

Figure 6.1 shows a practical list of steps to help improve the chances that the collection of data provides as accurate information as possible to fuel effective evaluation efforts. A top ten list offers a helpful tool for those planning to conduct surveys for evaluation and also for researchers who build on survey data to conduct their own analyses.

References

Adato, M. 2000. "The impact of PROGRESA on community social relationships". Final report. Washington, DC, United States: International Food Policy Research Institute.

Agar, M. 1980. *The Professional Stranger: An Informal Introduction to Ethnography.* New York, United States: Academic Press.

Ai, C., and E. Norton. 2003. "Interaction Terms in Logit and Probit Models". *Economics Letters* 80:123-129.

Anderson, M. 1986. "Cultural Concatenation of Deceit and Secrecy". In: R.W. Mitchell and N.S. Thompson, editors. *Deception: Perspectives on Human and Nonhuman Deceit.* New York, United States: State University of New York Press.

Angelucci, M., and G. De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1):486-508.

Antman, F. and D.J. McKenzie. 2007. "Earnings Mobility and Measurement Error: A Pseudo-Panel Approach". *Economic Development and Cultural Change.*

Aquilino, W.S. 1994. "Interview Mode Effects in Surveys of Drug and Alcohol Use". *Public Opinion Quarterly* 58:210-240.

Axinn, W.G. 1991. "The Influence of Interviewer Sex on Responses to Sensitive Questions in Nepal". *Social Science Research* 20:303-18.

Back, K.W. and J. Mayone Stycos. 1959. The Survey under Unusual Conditions: Methodological *Facets of the Jamaica Human Fertility Investigation.* Cornell University, Ithaca, New York, United States: Society for Applied Anthropology.

Bailar, B. 1976. "Some Sources of Error and their Effect on Census Statistics". *Demography* 13(2):273-286.

Barley, Nigel. 1983. *The Innocent Anthropologist: Notes from a Mud Hut.* London, United Kingdom: Penguin Books.

Becker, S., K. Feyisetan and P. Makinwa-Adebusoye. 1995. "The Effect of Sex of Interviewers on the Quality of Data in a Nigerian Family Planning Questionnaire". *Studies in Family Planning* 26:233-40.

Berk, M.L., N.A. Mathiowetz, E.P. Ward et al. (1987). "The Effect of Prepaid and Promised Incentives: Results of a Controlled Experiment". *Journal of Official Statistics* 3:449-457.

Bernard, H.R. 2000. *Social Research Methods: Qualitative and Quantitative Approaches.* Thousand Oaks, California, United States: Sage Publications.

Bignami-Van Assche, S., G. Reniers and A.A. Weinreb. 2003. "An assessment of the KDICP and MDICP data quality: interviewer effects, question reliability and sample attrition". *Demographic Research* SC1:31-75.

Black, D.A., M.C. Berger, and F.A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error". *Journal of the American Statistical Association* 95(451):739-748.

Blanc, A.K. and T.N. Croft. 1992. "The Effect of the Sex of Interviewer on Responses in Fertility Surveys: The Case of Ghana". Presented at the annual meeting of the Population Association of America, Apri1 30–May 1. Denver, Colorado: United States.

Bound, J., C. Brown, and N. Mathiowetz. 2001. "Measurement Error in Survey Data". In: J.J. Heckman and E. Leamer, editors. *Handbook of Econometrics.* Elsevier.

Bulmer, M., and D.P. Warwick. 1983/1993. *Social Research in Developing Countries: Surveys and Censuses in the Third World* (2nd. Edition). London, United Kingdom: UCL Press.

Caldwell, J.C., H.M. Choldin, L.F. Noe et al., 1970. *A Manual for Surveys of Fertility and Family Planning: Knowledge, Attitudes, and Practice.* New York, United States: The Population Council.

Cameron, A.C., and P.K. Trivedi. 2005. *Supplement to Microeconometrics: Methods and Applications.* Cambridge, United Kingdom: Cambridge University Press.

Casley, D.J. and D.A. Lury. 1981. *Data Collection in Developing Countries.* Oxford, United Kingdom: Clarendon Press.

Carroll, R., J. Ruppert, and L.A. Stefanski. 2006. 2nd Edition. *Measurement Error in Nonlinear Models.* London, United Kingdom: Chapman and Hall.

Church, A.H. 1993. "Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis". *Public Opinion Quarterly.* 57:62-79.

Cleland, J., and C. Scott. 1987. *The World Fertility Survey*. Oxford, United Kingdom: Oxford University Press.

Datta, A.R., M.W. Horrigan, and J.R. Walker. 2001. "Evaluation of a Monetary Incentive Payment Experiment in the National Longitudinal Survey of Youth, 1997 Cohort". Working Paper.

Deci, E.L. 1971. “Effects of Externally Mediated Rewards on Intrinsic Motivation”. *Journal of Personality and Social Psychology* 18(1):105-115.

Deming, W. 1944. “On Errors in Surveys”. *American Sociological Review* 19:359-69.

Feinberg, S.E. 1990. “Interactional Troubles in Face-to-Face Survey Interviews: Comment”. *Journal of the American Statistical Association* 85(409):241-244.

Ferber, R., and S. Sudman. 1974. “Effects of Compensation in Consumer Expenditure Studies”. *Annals of Economic and Social Measurement* 3:319-331.

Fowler, F.J., and T.W. Mangione. 1990. Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, California, United States: Sage Publications.

Fox, V. 2002. “Informe de Gobierno”. http://informe.presidencia.gob.mx/Informes/2002Fox2/website/cfm/index.cfm.

Frey, B.S. and F. Oberholzer-Gee. 1997. “The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out”. *American Economic Review* 87(4):746-755.

Gneezy, U. and A. Rustichini. 2000. “Pay Enough or Don’t Pay at All”. *Quarterly Journal of Economics* 115(3):791-810.

Green, R.T. and D.L. Alden. 1988. “Functional Equivalence in Cross-Cultural Consumer Behavior: Gift-Giving in Japan and the United States”. *Psychology and Marketing* 5(2):155-168.

Griliches, Z. and J. Hausman. 1986. “Errors in Variables in Panel Data”. *Journal of Econometrics* 31:93-118.

Grosh, M.E., and P. Glewwe. 1996. “Household Survey Data from Developing Countries: Progress and Prospects”. *The American Economic Review* 86(2):15-19.

Grosh, M.E. and J. Muñoz. 2000. “Metadata-Information about Each Interview and Questionnaire”. In: M. Grosh and P. Glewwe, editors. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Livings Standards Measurement Study.* Washington, DC, United States: World Bank.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York, United States: John Wiley.

Groves, R.M. 1991. “Measurement Error Across the Disciplines”. In: Biemer, P.P., R.M. Groves, L.E. Lyberg et al., editors. *Measurement Error in Surveys.* New York, United States: John Wiley & Sons, Inc.

Groves, R.M., R.B. Cialdini and M.P. Couper. 1992. "Understanding the Decision to Participate in a Survey". *Public Opinion Quarterly* 56:475-495.

Groves, R.M. and M.P. Couper. 1996. "Contact-Level Influences on Cooperation in Face-to-Face Surveys." *Journal of Official Statistics* 12(1):63-83.

Hausman, J. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left". *Journal of Economic Perspectives* 15(4):57-67.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error". *Econometrica* 47:153-161.

Hu, Y. and S.M. Schennach. 2008. "Instrumental Variable Treatment of Nonclassical Measurement Error Models". *Econometrica* 76(1):195-216.

Joseph, S. 1993. "Relationality and Ethnographic Subjectivity: Key Informants and the Construction of Personhood in Fieldwork". In: D.L. Wolf, editor. *Feminist Dilemmas in Fieldwork.* Boulder, Colorado, United States: Westview Press.

Kearl, B. (ed.) 1976. Field Data Collection in the Social Sciences: Experiences in Africa and the Middle East. New York, United States: Agricultural Development Council.

Krosnick, J.A. and D. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement". *Public Opinion Quarterly* 51:201-219.

Kruglanski, A., I. Friedman and G. Zeevi. (1971). "The Effect of Extrinsic Incentives on Some Qualitative Aspects of Task Performance". *Journal of Personality and Social Psychology* 39:608-617.

Leahey, E. 2008. "Methodological Memes and Mores: Toward a Sociology of Social Research". *Annual Review of Sociology* 34.

Little, R.J.A. and D.B. Rubin. 1987. *Statistical Analysis with Missing Data.* Hoboken, New Jersey, United States: John Wiley & Sons.

Massey, D.S. 1987. "The Ethnosurvey in Theory and Practice". *International Migration Review* 21:1498-1522.

Maynard, D.W., and N.C. Schaeffer. 2002. "Standardization and Its Discontents". In: D.W. Maynard, H. Hout-koop-Steenstra, N.C. Schaeffer et al., editors. *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview.* New York, United States: John Wiley and Sons.

Meijer, E., and T. Wansbeek. 2000. "Measurement Error in a Single Regressor". *Economics Letters* 69:277-284.

Mensch, B.S., P.C. Hewett and A. Erulkar. 2003. "The Reporting of Sensitive Behavior by Adolescents: A Methodological Experiment in Kenya". *Demography* 40(2):247-268.

Mensch, B.S., P.C. Hewett, H.E. Jones et al., 2008a. "Consistency in Women's Reports of Sensitive Behavior in an Interview Mode Experiment, Sáo Paulo, Brazil". *International Family Planning Perspectives* 34(4):169-176.

Mensch, B.S., P.C. Hewett, R. Gregory et al., 2008b. "Sexual Behavior and STI/HIV Status among Adolescents in Rural Malawi: An Evaluation of the Effect of Interview Mode on Reporting". *Studies in Family Planning* 39(4):321-334.

O'Muircheartaigh, C.A. 1982. Methodology of the Response Errors Project. Scientific Report No.28. London, United Kingdom: International Statistical Institute & World Fertility Survey.

O'Muircheartaigh, C.A., and A.M. Marckwardt. 1980. "An Assessment of the Reliability of WFS Data". Paper presented at the World Fertility Survey Conference, Methodology Session No.6. July 7-11. London: United Kingdom.

Plummer, M.L. et al., 2004. "'A Bit more Truthful': the Validity of Adolescent Sexual Behaviour Data Collected in Rural Northern Tanzania Using Five Methods". *Sexually Transmitted Infections* 80 (Suppl II): ii49-ii56.

Powell, B.A. and L. Pritzker. 1965. "Effects of Variation in Field Personnel on Census Results". *Demography* 2(1):8-32.

Presser, S. and L. Stinson. 1998. "Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance". *American Sociological Review* 63:137-145.

Rice, S.A. 1929. "Contagious Bias in the Interview: A Methodological Note". *American Journal of Sociology* 35:420-23.

Robinson, W.C. 1992. "Kenya Enters the Fertility Transition". *Population Studies* 46:445-457.

Sana, M., and A. Weinreb. 2009. "The Effects of Questionnaire Translation on Demographic Data and Analysis". *Population Research and Policy Review* 28(4):429-454.

Schaeffer, N.C., and S. Presser. 2003. "The Science of Asking Questions". *Annual Review of Sociology* 29:65-88.

Schaeffer, N.C., and E. Thomson. 1992. "The Discovery of Grounded Uncertainty: Developing Standardized Questions about Strength of Fertility Motivation". In: P. Marsden, editor.

Sociological Methodology 22:37-82. Washington, DC, United States: American Sociological Association.

Schennach, S.M. 2004. "Estimation of Nonlinear Models with Measurement Error." *Econometrica* 72(1):33-75.

Schober, M., and F.G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61:576-602.

Schwartz, B. 1967. "The Social Psychology of the Gift". *American Journal of Sociology* 73(1):1-11.

Simmel, G. 1950. *The Sociology of Georg Simmel.* Trans. and ed. K.H. Wolf. New York, United States: Free Press.

Singer, E., J. Van Hoewyk and M.P. Maher. 1998. "Does the Payment of Incentives Create Expectation Effects?" *Public Opinion Quarterly* 62(2):152-164.

Singer, E., J. Van Hoewyk and M.P. Maher. 2000. "Experiments with Incentives in Telephone Surveys". *Public Opinion Quarterly* 64:171-188.

Skoufias, E. 2005. "PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico". Research Report. Washington, DC, United States: International Food Policy Research Institute.

Skoufias, E., B. Davis and S. de la Vega. 2001. "Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA". *World Development* 29(10):1769-1784.

Skoufias, E., and B. McClafferty. 2001. "Is PROGRESA Working? A Summary of the Results of an Evaluation by IFPRI". Discussion Paper 118. Washington DC, United States: Food Consumption and Nutrition Division of the International Food Policy Research Institute.

Stecklov, G., P. Winters, M. Stampini et al., 2005. "Do Conditional Cash Transfers Influence Migration? A Study Using Experimental Data from the Mexican PROGRESA Program". *Demography* 42(4):769-790.

Stone, L., and J.G. Campbell. 1984. "The Use and Misuse of Surveys in International Development: An Experiment from Nepal". *Human Organization* 43:27-37.

Suchman, L., and B. Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews". *Journal of the American Statistical Association* 85(409):232-241.

Sudman, S., and N.M. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis.* Chicago, Illinois, United States: Aldine Publishing Company.

Titmuss, R.M. 1970. *The Gift Relationship.* London, United Kingdom: Allen and Unwin.

Tourangeau, R., L.J. Rips and K. Rasinski. 2000. *The Psychology of Survey Response.* Cambridge, United Kingdom: Cambridge University Press.

Tourangeau, R. and T.W. Smith. 1996. "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context". *Public Opinion Quarterly* 60:275-304.

Weinreb, A.A. 2006. "The Limitations of Stranger-interviewers in Rural Kenya". *American Sociological Review* 71(6):1014-1039.

Weinreb, A.A., S. Madhavan and P. Stern. 1998. "'The Gift Received has to be Repaid:' Respondents, Researchers and Gifting". Paper presented at the Annual Meeting of the Eastern Sociological Society, March 27. Philadelphia, Pennsylvania: United States.

Weinreb, A.A., and M. Sana. 2009a. "The Presence of a Third-party and its Effect on Survey Responses in Rural Malawi". Unpublished manuscript.

---------- 2009b "The Effects of Questionnaire Translation on Demographic Data and Analysis". *Population Research and Policy Review* 28(4):429-454.

Willimack, D.K., H. Schuman, B. Pennell et al., 1995. "Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey". *The Public Opinion Quarterly* 59(1):78-92.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross-Section and Panel Data.* Cambridge, Massachusetts, United States: The MIT Press.

Wooldridge, J.M. 2002. *Introductory Econometrics: A Modern Approach.* Mason, Ohio, United States: South-Western College Publishing.

Zagorsky, J.L. and P. Rhoton. 2008. "The Effects of Promised Monetary Incentives on Attrition in a Long-Term Panel Survey". *Public Opinion Quarterly* 72(3):502-513.