5 Outcomes and Instruments

This chapter covers planning for data collection. It gives practical advice for choosing what data to collect, selecting good indicators and instruments, and choosing the time, place, and frequency of data collection. The chapter has the following modules:

MODULE 5.1 Specifying Outcomes and Indicators

The programs we evaluate aim to improve the outcomes of participants, but to get the most out of an evaluation, it is useful to track outputs and intermediate outcomes as well as final outcomes. This module describes how developing a theory of change helps us map potential pathways of impact and ensure that we have indicators for each step of the process that are relevant to our particular context. It also shows that it is useful to include standard indicators that allow us to compare our findings to those of other studies in the literature.

We can use a theory of change to specify outcomes and indicators

Specifying good outcomes and the indicators we will use to measure them requires a deep understanding of the program being developed, the objectives of those implementing the program, and potential pathways through which the program or policy can impact lives, both

positively and negatively. Some of the terms we use are defined with examples in Table 5.1. A theory-of-change framework is a useful tool to use in systematically thinking through potential pathways to an impact and is best developed in close cooperation between those implementing and those evaluating a program. This theory of change is likely to lead us to develop indicators that are very specific to the context.

A *theory of change* is a structured approach used in the design and evaluation of social programs. It maps the logical chain of how program inputs achieve changes in outcomes through activities and outputs. We can go further and explicitly set down the assumptions that are needed to get from one step to another and the possible risks. For each step in the theory of change, we specify outcomes and indicators we will measure to help us understand whether the program has worked and also how it has worked. If the program is not successful, having indicators for these intermediate steps helps us understand at which step in the chain the program failed.

Indicators are observable signals of change

Outcomes are changes our program aims to achieve. Often these changes are conceptual, such as "economic empowerment of women"

TABLE 5.1 Data collection jargon

Term	Definition	Examples
Outcome	A change or impact caused by the program we are evaluating	Increase or decrease in women's empowerment, child health, corruption
Indicator	An observable signal used to measure outcomes	The number of women who spoke at a meeting, child arm circumference
Instrument	The tool we use to measure indicators	A survey question, achievement test, direct observation record
Variable	The numeric values of indicators	Self-evident
Respondent	The person or group of people we interview, test, or observe to measure the indicators	Individuals; their teachers, colleagues, or family

or "improved learning." *Indicators* are observable signals of these changes, such as "monthly income of women in a village" or "ability to read a simple paragraph." Indicators measure whether the changes the program is designed to bring about are in fact happening. Test scores, for example, can be an indicator of the more abstract concept of learning. For an indicator to be any good, there needs to be a strong logical link connecting it to the relevant outcome we are ultimately trying to measure.

The logical validity of the indicator depends on the context

Not only must the indicator logically follow from the outcome concept, but this logic also has to be valid for the context.

Say our outcome of interest is teacher effort. We need to choose indicators that measure how our program is affecting the efforts of teachers. In rural areas with remote schools, teachers face high travel costs. In a context of high levels of teacher absenteeism, whether teachers show up at school could be a good indicator of teacher effort. However, teacher attendance as an indicator of effort would not work as well in urban areas where teachers face very low travel costs and shirking may instead take the form of socializing with colleagues in the staff room rather than teaching. In this context, better indicators of effort might be "time spent with students after school," "time spent teaching instead of socializing in the staff room," or "frequency of assigning homework."

Module 5.3 discusses in more detail how to assess the validity of indicators in a given context.

Specifying outcomes from the literature

If an evaluation is to be influential in broader policy debates, it is useful if its results can be compared with those of other leading studies in the field. To facilitate comparison, it is helpful to include outcome measures and indicators that are commonly used in the literature for assessing programs with similar objectives.

Take for example the goal of increasing the quantity of education. Some studies only look at enrollment rates of students, but an increasing number of studies also examine attendance rates by counting how many children are at school during unannounced visits. Collecting attendance data allows an evaluator to calculate the total increase in schooling generated by a program (measured in weeks or years of additional schooling). For a policymaker choosing between alterna-

tive approaches to increasing schooling quantity, it is useful to be able to see which of many approaches is the most cost-effective, but this is possible only if outcomes are defined in the same way across projects. (Module 9.3 discusses ways to calculate and use comparative cost-effectiveness analyses.)

Consensus on the best outcome measures and indicators is always evolving. Examining the recent literature for the most current measurement standards allows us to benefit from what others have learned about assessing outcomes reliably and accurately. This is particularly important in social and economic evaluations whose content overlaps with that of other disciplines. For example, a microfinance program aiming to empower women will incorporate frameworks from the economics, sociology, and psychology literatures; a health program will be influenced by the medical and public health literatures and an education program by the education and cognitive science literatures; and a political empowerment project can incorporate aspects of the political science and psychology literatures. These literatures are a deep source of useful outcomes and indicators, and many have their own conventional outcome measures. For example, in cognitive development there is a battery of standardized tests; for learning there are internationally standardized achievement tests; and for nutrition there are measures of arm circumference, weight for age, and weight for height.

Case studies on specifying outcomes and indicators

The following case studies illustrate a two-step process for selecting outcomes and indicators during program evaluation: (1) mapping the theory of change and (2) determining indicators that are the logical consequences of each step in the theory of change, when possible drawing on the existing literature. We also give examples of specifying assumptions needed to move through the theory of change.

Example: An HIV education program in Kenya

A program in Kenya provided in-service training for teachers to improve their delivery of HIV prevention education in primary schools. The training focused on how to teach HIV/AIDS prevention best practices while teaching other subjects and how to start and run after-school student health clubs devoted to HIV/AIDS education.[1]

1. This study by Esther Duflo, Pascaline Dupas, and Michael Kremer is summarized as Evaluation 4 in the appendix.

A simple theory of change for the program says that (1) teacher training (2) increased HIV education, which (3) increased knowledge of prevention best practices, which (4) reduced unsafe sexual behavior, which (4) led to reduced HIV infection rates. The theory of change links five steps: teacher training, HIV education, knowledge of best practices, sexual behavior, and incidence of HIV infection. Although the policy variable of interest is HIV status, the program directly targets only the intermediate outcome of unprotected sexual behavior and knowledge.

For each of the concepts in the causal chain, we need to find a concrete indicator in the real world that we can observe and measure (Table 5.2).

HIV status is the ultimate outcome of interest. If we can measure HIV status in both our treatment and our comparison groups, we can know whether our program changed HIV infection rates. Although testing HIV status is common in the literature, it may be too expensive as well as ethically, politically, or logistically infeasible in our context. If this is the case, we can use other indicators as proxies for HIV rates. In the literature it is common to use a broader range of sexually transmitted infections (STIs) as a signal of unprotected sex. In this case, our observable indicators would include biomarkers for a range of STIs common in this context, such as herpes, syphilis, gonorrhea, chlamydia, hepatitis, and human papillomavirus. The results of tests for these STIs are commonly used in the literature as indicators of risky sexual behavior. Childbearing would also be a good proxy for STIs and HIV infection, because the same unsafe sexual behavior that leads to childbearing can also lead to STIs and to HIV infection.

We may also be interested in more proximate outcomes, such as changes in knowledge, that are not captured by HIV status. Indeed, HIV status is not a valid measure of *knowledge* of best practices, because someone may know the best prevention methods but choose to not practice them. Measuring these intermediate outputs and outcomes would allow us to learn about the underlying mechanisms that can translate program inputs into impacts. For example, if the program failed to achieve the desired impact, did it fail because it did not change knowledge or because changing knowledge did not change behavior?

Example: Achieving political reservations for women in local politics

India amended its federal constitution in 1992, devolving the power to plan and implement development programs from the states to rural

TABLE 5.2 Logical framework for HIV education program

Inputs, outputs/outcomes	Objectives hierarchy	Indicators	Assumptions or threats
Inputs (activities)	Teachers are trained to provide HIV education.	Hours of training implemented	Teachers engage with the training and learn new concepts.
Outputs	Teachers increase and improve HIV education.	Hours of HIV education given; application of the program's teaching methods	Teachers are receptive to changing their HIV education approach, and outside pressures do not prevent them from implementing a new curriculum.
Outcome (project objective)	Students learn prevention best practices.	Test scores on HIV education exam	Students engage with the new curriculum and understand and retain the new concepts.
Impact (goal, overall objective)	Students reduce their unprotected sexual behavior; incidence of HIV infection decreases.	Self-reported sexual behavior; number of girls who have started childbearing; HIV status	The information students learn changes their beliefs and their behavior. Students can make decisions and act on their preferences for engaging in sexual behavior.

village councils. Councils now choose what development programs to undertake and how much of the village budgets to invest in them. The states are also required to reserve one-third of council seats and council chairperson (*pradhan*) positions for women. Most states developed lotteries so they could use random assignment to create a schedule as to which third of the villages are required to reserve the council chair position for a women in a given election cycle.[2] The lotteries created an opportunity to assess the impact of reservations on politics and government by answering the following questions: Do policies differ when there are more women in government? Do the policies chosen by women in power reflect the policy priorities of women?

A basic theory of change for the evaluation was (1) legislation mandating quotas for women leaders is passed, (2) more women are made chairs of their village council, (3) investment decisions better reflect the preferences of women, and (4) the quality of public goods preferred by women increases. In this case there were relatively few standard indicators in the literature for the researchers to rely on, and most of the indicators were specific to the evaluation context.

Legislation mandating quotas for women leaders is passed Even after quotas for women was mandated by the Supreme Court, the necessary legislation had to be passed in each state. This input could be confirmed through legislative records.

More women are made chairs of their village councils Once the legislation was passed, it still had to be implemented at the village level. Implementation could be checked by determining the gender of council chairs in office from official records.

Women's participation in village councils increases To ensure that the council investments reflect the community's priorities, the amendment included two requirements that allow members of the community to articulate their priorities. First, the councils must hold a general assembly every six months or every year to report on activities in the preceding period and must submit the proposed budget to the community for ratification. Second, the council chairs must set up regular office hours to allow people to submit requests and complaints.

2. Two studies on this policy, one by Raghabendra Chattopadhyay and Esther Duflo and the other by Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova, are summarized as Evaluation 12 in the appendix.

Having female council leaders may embolden women to voice their policy preferences at general assemblies, and being able to direct their queries to other women could further encourage these women to express their views. If women become empowered, they will take advantage of available political channels. Our observable indicators of this behavior are women's attending and speaking at general assembly meetings and their submitting requests and complaints.

Investment in public goods better reflects the preferences of women The councils decide which development programs to implement and how much to invest in them. They have to choose the programs from a list of development areas, including welfare services (services for widows, care for the elderly, maternity care, antenatal care, child healthcare) and public works (providing drinking water, roads, housing, community buildings, electricity, irrigation, and education).

We can rely on public records for estimates of or expenditures on public goods and services. But we also want to verify that the officially recorded public goods and services were actually delivered on the ground. We can do this by taking an inventory of public goods of different types in the community and asking when they were built and when they were last repaired.

We also need to determine which public goods are preferred by women. It takes time and potentially courage to speak up during a general assembly meeting or to submit a service request or complaint. Speaking up about one's priorities can also have real consequences, such as affecting the sectors or projects in which public money is invested. This means that people are unlikely to speak unless they have a compelling interest in the issue at stake. Thus an observable indicator of people's true preferences is the public goods they address in their requests, complaints, and speeches to the general assembly. Sorting the issues raised by gender should help us determine the policy preferences of men and women. It is worth noting that we would not expect to see a change in investment in a good that was highly valued by women but also highly valued by men. In this case there would be investment in both quota villages and nonquota villages, so the introduction of quotas would not change the pattern of investment.

The quality of public goods preferred by women increases Quotas for women would not benefit women if they led only to increases in invest-

ment in goods preferred by women but not to improvements in the quality or quantity of these goods. We can measure the quality of goods. In the case of water quality, for example, we can test for the presence of microbes in the water.

Assumptions, conditions, and threats

Our theory of change rests on a number of assumptions and conditions that must be met for the chain of events in the theory of change to hold. In the logical framework for our evaluation of the program to increase women's participation in village councils (Table 5.3), we map these assumptions:

- *Reservations for women were implemented properly.* The amendment directed the states to (1) establish the rural councils, (2) devolve all powers to plan and implement rural development programs to the councils, (3) ensure that council elections are held every five years, and (4) ensure that one-third of all council seats and council chair positions are reserved for women.
- *Seeing women leaders encourages other women to participate more fully.* When quota legislation was passed there was concern that women leaders would have power in name only. In this scenario there would be little reason for other women to become more active participants in village councils, to speak up, or to voice complaints.
- *Women have different preferences from those of men.* This implies that if we sort preferences by gender, we will see a difference. For this we need to catalog the content of queries made by type (water, roads, agriculture, etc.) and check whether men and women have different preferences and in which areas.
- *Some democracy exists.* There should be elections every five years, with councilors popularly elected to represent each ward. The councilors should elect from among themselves a council chairperson. Decisions should be made by a majority vote, and the chairperson should have no veto power. That there is democracy implies that investments would reflect the preferences of the constituents.
- *The democracy is imperfect.* If there were perfect democracy, elected officials would perfectly channel the wishes of all their constituents, and the leader's gender and preferences would not

TABLE 5.3 Potential indicators of the outcomes of a policy to increase women's participation in village councils

Inputs, outputs/outcomes	Indicator	Assumptions and threats
Inputs		
Quotas for women are passed.	Passage of legislation in state legislatures	Supreme court mandate is translated into effective legislation at state level.
Outputs		
There are more women leaders.	Number of women leaders in council chair positions	Quota legislation is implemented as designed in villages.
Political participation increases.	Number of complaints brought by women	Seeing women leaders encourages women to voice complaints.
	Number of women speaking at general meetings	Seeing women leaders emboldens women to speak up at meetings.
Impact		
Public goods investments more closely match women's priorities.	Types of public goods mentioned in women's queries versus men's queries	Women's preferences for public goods differ from men's.
	Number of public goods of different types	The system is democratic enough to respond to an increase in women's queries and public statements but not democratic enough to have taken their views into account prior to having women leaders.
	Repairs to public goods of different types	Responses to political pressure from women will impact repairs as well as new investments.
	Recently built public goods by type	New investments will be more in line with women's needs than were older investments.
The quality of public goods that are priorities for women improves.	Reduced presence of microbes in drinking water	Greater investment in areas of priority to women will translate into better-quality services.

matter. But the chairperson is the only councilor with a full-time appointment and so wields effective power. That the democracy is imperfect implies that investments would reflect the preferences of the leader.

- *Increases in investment translate into more public goods of better quality.* Political pressure from women could lead to more investment in the goods that they prefer, but that does not necessarily translate into more or better-quality goods if the money is spent poorly. Some commentators were concerned that because women leaders were inexperienced the investments would not translate into real gains. If indicators such as repairs of goods preferred by women were to be good measures, the investments would have to take the form of repairs as well as new building.

MODULE 5.2 Specifying Data Sources

Once we have specified our outcomes of interest and selected the indicators we plan to use to measure those outcomes, we need to specify how that data will be collected. This module discusses how we go about choosing how we will collect the data (whether to use existing data or collect our own through survey or nonsurvey instruments), when we will collect the data, who will perform the data collection and from whom they will collect it, and where they will collect it.

Finding existing data sources or collecting new data

We must decide between using existing administrative data sources and collecting our own data through surveys or nonsurvey instruments.

Using administrative data

How will we acquire data on our outcomes of interest? In some cases, existing administrative data can be a good source. Administrative data are records collected by governments or civil society organizations, usually in the context of program administration.

First, we must evaluate whether the administrative data will be suitable for answering our research questions. The following issues should be considered:

- Do the data actually exist?
- Have the data been consistently collected?

- Does the data set cover our population of interest?
- Do the data cover the outcomes that interest us?
- Are the data reliable and unlikely to have been manipulated?

Limitations Administrative data tend to come in two kinds: basic data that are collected on everyone (for example, results on school completion exams) and more detailed data that are collected on a random sample of individuals (most government statistics come from representative surveys). Typically, administrative data that are collected on all individuals are not very detailed and may not be sufficient to answer all the questions we have. There are exceptions: for example, medical records contain detailed information about the treatment and outcomes of all patients, and prison records can contain quite a lot of information on prisoners.

Large-scale national surveys tend to collect very detailed data but only for a subset of the population. For example, the Demographic and Health Surveys collect rich data from individuals sampled to be nationally representative. In most cases, however, not enough people from our treatment and comparison groups will have been sampled for the survey to give us sufficient sample size for the analysis.

Example: Electoral outcomes of politicians In Brazil a policy randomly assigned the timing of corruption audits of municipalities. Researchers were then able to use existing administrative data on electoral outcomes to evaluate the impact on the electoral outcomes of incumbent mayors of making information on corruption available.

For further reading

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123 (2): 703–745.

J-PAL Policy Briefcase. 2011. "Exposing Corrupt Politicians." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/exposing-corrupt-politicians.

Example: Collecting overdue taxes in the United Kingdom The Behavioral Insights Team in the Cabinet Office of the United Kingdom tested different ways to encourage taxpayers to pay off their debts to the government. Roughly 70 percent of those in arrears paid up after receiving a letter from the authorities, but the government wanted to see if

they could increase this rate, because taking further steps toward litigation is expensive. They developed alternative ways to draft the letter requesting payment, drawing on lessons from behavioral research. They then randomized which debtor was sent which letter. Outcomes were measured using tax administration records.

For further reading

Cabinet Office Behavioral Insight Team. 2012. "Applying Behavioral Insights to Reduce Fraud, Error, and Debt." Accessed May 28, 2013. https://update.cabinetoffice.gov.uk/sites/default/files/resources/BIT_FraudErrorDebt_accessible.pdf.

Example: Test scores of high school graduates Researchers evaluated the impact of a policy that randomly selected families to receive vouchers for private schooling. Later the researchers examined administrative records on registration and test scores from a government college entrance examination. The test scores also gave the researchers a good proxy for high school graduation, since 90 percent of all graduating high school seniors take the exam.

For further reading

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3): 847–862.

This study is summarized as Evaluation 10 in the appendix.

Collecting our own data

In many cases we cannot rely only on data collected by others to measure our outcomes because the data do not exist, they are unreliable, or the sample size is insufficient. Collecting our own data allows us to create a custom data set on exactly those outcomes that are most relevant to our research questions. However, collecting data is expensive, and we will face many trade-offs in deciding exactly what data to collect and when.

Surveys The most common method of collecting data for impact evaluation is to administer surveys. Surveys are good for collecting data on lots of different questions from the individuals in our study. They are relatively quick, and (compared to some nonsurvey instruments) relatively inexpensive. However, for some outcomes we may

worry that the information people report in surveys is unreliable because they may have forgotten what happened in the past or they may misreport potentially sensitive information.

Nonsurvey instruments Although using surveys is the most common way to collect data, there are also a host of nonsurvey instruments that can help us quantify difficult-to-measure outcomes. These range from simple vignettes an enumerator presents to respondents, often as part of a survey (discussed in greater detail later in this chapter), to the collection of complex biomarker data. These instruments enable us to rigorously quantify activities or outcomes that participants would not report in a survey (such as corruption or discrimination), that they may not know (such as whether they have an HIV infection or subconscious racial or gender bias), or that need more nuanced measurement (for instance, small changes in health outcomes such as levels of anemia).

Nonsurvey instruments also have important limitations, which vary by instrument. In Module 5.4 we discuss a number of nonsurvey instruments in detail, noting the advantages and limitations of each. Many nonsurvey instruments, however, are quite expensive for the number of outcomes they measure compared to a survey that can cover many outcomes in one survey.

Specifying the subjects and respondents

Who will be our subjects and respondents? This question encompasses many others: Who is a representative unit of observation? Who is the respondent that knows the information we need about the unit of observation? Who is the respondent that will give the most reliable information? Who is the efficient respondent that will allow us to gather the most information at one time?

Who will be our subjects and respondents?

Who is subject to the treatment? The person whose outcomes we will measure depends on the question we are trying to answer. For example, if we want to know whether a business training program increased knowledge of business practices, we will test the participants. But the choice is not always so straightforward.

If we provide a clean water program and want to study its health effects, we can look at the health of everyone in a family because they are all likely to suffer from diarrhea if they drink unclean water. How-

ever, it probably makes sense to focus only on diarrhea among children under age 5 for two reasons: (1) children under that age are much more likely to die of waterborne diseases, and thus we care more about diarrhea in this age group, and (2) the incidence of waterborne diseases is higher in this age group, so it will be easier to pick up an effect in this age group.

In contrast, for microfinance programs we may want to look not only at the person who got a loan (if the MFI lends to women only, the female head of the household) but also at her spouse. Maybe we would see no effect if we looked only at spending on the female head but would see a change in total household consumption or in the size of a business owned by the male head.

The question of whom to use as respondents in collecting data is less obvious when we are dealing with aggregates such as households. Whom do we include in the aggregate? How, for example, do we define a household when dealing with nonnuclear families or polygamous families?

Who is representative? In many cases we do not measure every person in the treatment and comparison group, and in these cases we need to choose whom to sample. For example, imagine that we are doing a communitywide iron supplementation program. We decide that it would cost too much to take blood samples from everyone in both the treatment and the comparison communities. A key consideration then becomes from whom we will take blood so that our subsample is representative.

One common sampling methodology is to go to the center of the community and interview someone in every second (or third) house to which the interviewer goes. Another is to go to the center of a community and start talking to passersby. But the people we meet in the middle of the day in the middle of a community are not random. In many countries richer people tend to live near the center of a community and many people will be out working in the middle of the day, so it is a very special group of people we meet in the middle of the day. The only way to be sure that we get a truly representative group of people for our data collection is to get a listing of everyone who lives in an area and then randomly select people or households to be measured.

Who knows the information we need? We want to interview respondents who have the information that we are looking for. If, for example,

we are collecting information on child health—say, the number of cases of diarrhea in the past week—we want to go to the primary caregiver. And we want to make sure that we obtain comparable responses; in other words, if we decide to interview the mother because she is the primary caregiver, we should make sure to systematically interview mothers and not substitute aunts or sisters for convenience. Allowing no substitutions can be costly. If enumerators arrive in a household and the mother is absent, they might have to come back another day to interview her.

In contrast, if we want information on the number of public works projects in a village over the past year, we can do a participatory appraisal in which we simultaneously interview a group of, say, 20 villagers to get an aggregate picture of these projects. This will give us information that is more accurate than that we would get from interviewing one person, no matter how well chosen.

Who is unlikely to manipulate the information? Which respondent has the least incentive to manipulate the information? Some questions are sensitive, and respondents may have an incentive not to be straightforward in their answers, in which case we need to think carefully about the incentives of different potential respondents.

Consider teacher attendance. Both students and teachers can supply this information. But the teacher has a stronger incentive to misreport her attendance because she may fear sanctions from supervisors, or she may misreport simply because she does not like to think of herself as shirking. If we cannot directly observe attendance and have to rely on reported information, the attendance reported by the students may be closer to the truth than that reported by the teacher.

Who will be most efficient in reporting data? Efficiency has to do with the amount of information we can get from one person in one sitting. If, for example, we want information on child health and education and our indicator for the former is test scores on report cards, the parents may be a more efficient source than the teachers. From the teachers we will get only test scores, but from the parents we can get both the test scores and the health information.

Who will do the measuring?

It is important to think carefully about whom we will choose to be the enumerators who collect the data.

The same people must interview both the treatment and the comparison groups. It can seem expedient to have program staff who are already working with the treatment group interview them and hire others to interview the comparison group. However, it is critical that the same approach be used and the same team interview both the treatment and the comparison group. Otherwise it is possible that differences in the data will be due to differences in interviewing style, procedure, or social interaction—not the program.

Enumerators should not be program staff or anyone the participants know from the program It can be tempting to try to save on costs and human resources by using program staff to conduct the survey with both treatment and comparison groups. However, the treatment group may feel less free to answer questions honestly if the enumerator is someone they recognize from a program that provided services. Also, it is important to have well-trained enumerators who are skilled in administering surveys and know how to avoid asking leading questions. We cannot simply hand a survey (even a well-designed one) to general staff and hope they can administer it well.

The characteristics of the enumerator can matter We need to think carefully about the social interaction between the enumerator and the respondent. In some cases, for example, it is best if the enumerator is the same gender as the respondent. Minimizing language barriers is also very important: ideally, enumerators will be fluent in all the local languages spoken in the area in which they conduct surveys. When enumerators work in small teams, at least one enumerator should be fluent in each relevant language.

We should have a plan to check for cheating and announce it in advance We need to have a carefully made plan in place to check for cheating or carelessness by enumerators. Enumerator shirking could take a number of forms. Enumerators may save time by filling in surveys without ever interviewing the respondents. But an enumerator can also cheat by not being careful to track the correct respondent and simply substituting another person who happens to be present for the person in our sample. Or she may save time by taking advantage of skip patterns in the survey. For example, often there are certain questions in a survey on which a "yes" answer will lead to many more questions than a "no." If the enumerator writes down a lower number

of household members in response to an early question, she will have fewer people about whom to ask questions. Similar problems could arise in response to questions about the number of crops grown, whether a person has ever taken a loan, or whether the household has experienced episodes of illness. Cheating and carelessness of this kind can invalidate the entire evaluation.

Different researchers have different approaches to checking for cheating, but the most common forms are performing random back-checks and looking for patterns in the data. In random back-checks, a separate team of enumerators randomly selects people in the sample, whom they then visit and administer a short survey. The survey asks whether someone recently visited them to ask some questions and repeats a few objective questions from the survey to which the answers are unlikely to change over time. Back-checking is most effective when enumerators are warned in advance that their work will be checked in this way and when every enumeration team is back-checked in the first few days of data collection so the threat is credible.

One rule of thumb is to administer back-check questionnaires to approximately 10 percent of the people surveyed. For larger surveys, slightly fewer than 10 percent may suffice, but for small surveys about 15 percent may be needed. The important guiding principle when deciding how many back-checks to do is to make sure that every team and every surveyor is back-checked as soon as possible. One option to consider is to have a more aggressive back-check process in the first few weeks of the survey and then reduce it to 10 percent of surveys through the rest of the surveying phase.

We also need to check for patterns in the data that suggest cheating. For example, the management team for the data collection process can check in real time for a high number of "no" answers to questions that let enumerators skip forward in the survey. Providing enumerators with clear definitions can also be helpful. For example, because there is an incentive to shrink household size so that the enumerator has fewer people about whom to ask questions, we must clearly define what we mean by a household. This will be culturally dependent, but a common definition is people who eat together.

Finally, computer-assisted interviewing using such things as GPS-enabled devices can make it easier to monitor for some kinds of cheating. Requiring enumerators to fill in the GPS locations of different households or interview sites makes it possible to check whether the enumerators did in fact visit those locations rather than filling in all

the surveys at one location. If data are collected electronically and quickly downloaded onto a main server, it also becomes much easier to spot suspicious patterns in the data early in the data collection process.

Specifying the time and frequency

We need to decide when and how often to collect data. Should we conduct a baseline survey before the program is rolled out? Should we collect data throughout program implementation? How long do we wait before doing an endline survey? How often should we visit, and how long should these visits be?

Should we conduct a baseline survey?

A *baseline* survey could be worthwhile when (1) the sample size is limited, (2) the outcomes are specific to individuals, (3) we want to show that our treatment and comparison groups were balanced at the start of the evaluation or (preferably) ensure they are balanced by doing a stratified randomization, (4) we want to analyze the data by subgroup or include control variables in the final analysis.

Limited sample size Collecting baseline data can help us increase our statistical power, which is particularly important when our sample size is small. In other words, for a given sample size we will be able to detect smaller program effects with baseline data than otherwise. In our analysis we can use baseline data to reduce unexplained variation in the outcomes of interest. For example, children vary widely in their test scores. If we have a test score for one child both before the program starts and after it ends, our statistical analysis can use those baseline data to make our estimate of the impact much more precise. However, a baseline can be expensive, so we will want to trade off the cost of doing a baseline with the cost of increasing our sample size. (For details, see Chapter 6 on statistical power.)

Individual-specific outcomes A baseline is particularly useful when the outcomes are person specific. For example, cognitive abilities, test scores, and beliefs tend to be highly correlated over time for a given individual. If we collect baseline and endline data on the same individuals for these variables, we will be able to explain a lot of the variance between individuals. For other variables, such as agricultural yields, there is much lower correlation over time for the same person.

In such a case, having a baseline will provide less additional statistical power.

Balance Sometimes the luck of the draw means that our randomized assignment results in unbalanced treatment and comparison groups; for example, the treatment schools may have a disproportionate number of the more educated teachers. Thus, if we find that children in treatment schools have higher test scores than those in the comparison schools, it may be due to the program or it may simply be due to the imbalance of educated teachers. Imbalance is more likely when the sample is small, and it makes interpreting the results more difficult.

Having baseline data can be reassuring if it confirms that our randomization led to good balance between treatment and comparison groups. This balance can also help confirm that randomization was in fact carried out (for example, if we are relying on a randomization done by others). However, a baseline can also reveal that our groups are not balanced, and there is little we can do if this is the case. A better strategy (as discussed in Chapter 4) is to do a stratified randomization. Baselines are useful for generating the data on which to stratify the randomization. This approach does require that the baseline data be collected, entered, and cleaned before randomization takes place and before the program is rolled out. If we are taking this approach, it is important to explain to those implementing the program that there will need to be a delay between the baseline and the rollout of the program.

Subgroup analysis and baseline controls Baseline data allow us to define subgroups for analysis. For example, if we have income data, we can categorize people by income brackets and then check how the effects of the program vary by income. Or if we have pretest scores, we can check whether the program benefited all students or just those who, based on pretest scores, were already doing well. It is not possible to do this if we have data on outcomes only after the program was implemented because we don't know who was poor or doing badly before the program began. Analyzing effects by subgroup often helps us answer important policy questions.

Similarly, baseline data allow us to include baseline characteristics such as income as control variables in our final analysis, which can help increase our statistical power (see Modules 6.4 and 8.2).

When should we start data collection after the program rollout?

How long after the program rollout should the follow-up data collection begin? That depends on whether there are novelty effects and on the lag between the program's inception and its impact.

Novelty effects arise when some aspect of a new program excites the participants and artificially raises outcomes in the short term. For example, if we monitor teacher attendance with cameras, the cameras themselves may cause excitement and temporarily increase the attendance of both students and teachers. It may be better to measure impact when this novelty has worn off in order to get a better estimate of the medium-term impact of monitoring.

Program impacts are rarely instantaneous. Impact should be measured after the program has had time to achieve its effects. For example, even if a program reduces teacher absenteeism instantaneously, the effect of more regular teaching may take time to show up in learning and test scores.

How frequently should we collect data?

The more frequent our data collection, the more likely we are to capture intermediate outcomes and patterns in the outcomes.

A benefit of frequent collection is that it can enrich the explanation of program impacts. A study in Kenya evaluated a program in which teachers were paid based on attendance, with school principals responsible for recording absences and awarding regular attendance bonuses.[3] The researchers did random spot checks three times a month to measure attendance independently. The resulting data allowed them to cross-check the attendance data they received from the principals and also to see the pattern of attendance over the course of the program. They found that the attendance level was high at the beginning, but after six months it deteriorated to its normal level again, and that change coincided with an increase in the number of absences marked as presences. This analysis was possible only because the researchers collected intermediate outcomes frequently. If they had measured attendance only at the end of the evaluation period, that would have sufficed to estimate the program's impact, but they would

3. Michael Kremer and Daniel Chen, "An Interim Report on a Teacher Attendance Incentive Program in Kenya," Harvard University, Cambridge, MA, 2001, mimeo.

not have seen the full story of the initial attendance gains that were later undermined by principal discretion.

A downside of frequent data collection is the cost, not just to the evaluation but also to our respondents, who have to admit enumerators to their homes or classrooms and fill out the surveys. In deciding on frequency, we face a trade-off between detail and cost.

Sometimes too-frequent surveying can itself become an intervention and change outcomes. Another evaluation in Kenya estimated the effect on child health and diarrhea of distributing chlorine for household water purification. In the treatment and comparison groups, the researchers varied how frequently they visited families to ask them about their use of chlorine in their drinking water. They found that visiting the families frequently caused them to increase their use of chlorine. In a sense, the data collection acted as a reminder to chlorinate the water. The comparison group with frequent data collection no longer provided a representation of chlorine use in the absence of the program.[4]

When should we end the data collection?

When should we do the endline survey? When the program ends? After the program ends? And if we want to know the long-term effects, how long after the program ends should we wait? The decision is usually a trade-off between waiting to get longer-term results and getting a result in time for the findings to inform decisionmaking. Attrition is also an important factor for long-term follow up. The longer we wait to do the endline survey, the higher the attrition rate is likely to be, because people will move and even die over time.

How does the randomization design influence the timing of data collection? The timing of data collection may vary with the randomization design, particularly for the comparison group.

In an *encouragement design* it may take a while for those who are encouraged to take up the program. If we wait too long to collect data, however, even those who have not been encouraged may take up the program. We need to do the endline survey at a point when take-up is high in the encouraged group but not yet high in the group not encouraged.

4. Alix Peterson Zwane et al., "Being Surveyed Can Change Later Behavior and Related Parameter Estimates," *Proceedings of the National Academy of Sciences USA* 108 (2011): 1821–1826.

In a *phase-in design* we need to conduct the endline survey before the last group receives the program. However, in some versions of a phase-in design it is still possible to measure long-term results. The comparison for the long-term effects is between those who had the program for a longer period and those who had it for a shorter period of time. For example, see the evaluations of the school-based deworming program in Kenya.[5]

In a *rotation design* we must collect data each time the treatment rotates between the two groups.

Specifying locations: What locations make for efficiency?

Imagine that we are running an after-school book club to see whether it helps increase reading comprehension. We randomize some schools to receive the club. We could test children on reading comprehension at school or at home. It will be much more efficient to test them at school because most of them will be gathered in one place at one time. The schools (even the comparison schools) may be happy to cooperate because they will receive feedback on their students. We may still have to do some tests at home to follow up on students who have dropped out of school or are absent when we do the test, but this will still be less expensive than doing all the tests at school.

MODULE 5.3 Assessing and Field Testing Outcome Measures

Now that we have a list of indicators, how do we decide which are best for our evaluation? This module discusses the need to look for four qualities in an indicator: it must be logically valid, measurable (observable, feasible, and detectable), precise (exhaustive and exclusive), and reliable.

Criteria for a desktop assessment of potential indicators

We will have a long list of potential indicators after we review the literature and map outcomes and indicators to the program's theory of change. This list may be too long. It would be too costly, not to mention confusing, to measure every possible variable that could be

5. Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel, "Worms at Work: Long-Run Impacts of Child Health Gains," working paper, Harvard University, Cambridge, MA, October 2011. http://www.economics.harvard.edu/faculty/kremer/files/KLPS-Labor_2012-03-23_clean.pdf. See also Evaluation 1 in the Appendix, which summarizes this study.

changed by the program. Ultimately, we need to test our indicators by piloting them in the field. Before doing that we can conduct a "desktop assessment" of our potential indicators. We can compare them according to four qualities of a good indicator to winnow our list of potential indicators to the most promising ones.

Logically valid

In the specific context of the program we evaluate, there must be a logical link connecting our outcome of interest at the conceptual level and the indicator we observe and measure.

Example 1: HIV status is a logical indicator of unsafe sexual behavior HIV status is a valid indicator of unsafe sexual behavior. The logical link is as follows: unsafe behavior can lead to infection with HIV, infection causes the body to produce HIV antibodies after a certain period, and an HIV test determines whether these antibodies are present in the blood.

*Example 2: HIV status is not a logical indicator of knowledge of safe sexual practice*s HIV status is not a valid indicator of knowledge of safe sexual practices; someone may know the best prevention methods but not practice them.

Example 3: Childbearing is a logical indicator of unsafe sexual behavior Unprotected sex can lead to pregnancy, which can lead to the birth of a child. Thus childbearing is a valid measure of unsafe behavior.

We must also consider whether there is any potential for the program to change the logical relationship between an indicator and the outcome we are trying to measure. An indicator could be correlated with our outcome of interest, but if the intervention changes that correlation, it is no longer a good indicator. For example, having a sheet metal roof is a good proxy for wealth in many communities in developing countries, so we often ask about the roofing material of a house to understand wealth levels. But if our intervention involves supporting house renovation and we see more corrugated iron roofs in our treatment group than in our comparison group, we cannot necessarily assume that general wealth has increased as a result of the program.

Measurable

A measurable indicator is observable, feasible, and detectable.

Observable An indicator must be a behavior or state that can be observed in the real world. Thus happiness is not an indicator, but laughter or self-reported happiness could be. Learning is not an indicator, but being able to read and do arithmetic are. This is an important distinction to maintain. We are not done defining our outcomes until we have identified some indicators that are observable.

Feasible Indicators must be feasible to measure—politically, ethically, and financially. Being feasible is context specific. Thus HIV status is observable in theory, but it may not be a good indicator for a school-based HIV education program if it is infeasible to test children for HIV. If certain questions are too sensitive to ask, indicators derived from those questions are not feasible.

Detectable Indicators must be detectable with the instruments and statistical power of our specific experiment. Thus infant mortality (number of deaths of children aged 12 months or younger per 1,000 live births) might not be a good indicator for a maternal and child health program. If infant mortality occurs relatively infrequently, even if the program has an effect there may not be enough of a change in the numbers of deaths to detect the effect. If we are testing a program to improve water quality, we might want to measure the number and severity of diarrheal episodes as a more detectable alternative.

In our HIV education example we are targeting children in the last grade of primary school, age 15 years on average. The government does not authorize testing of primary school children for HIV, nor do we have the infrastructure to test for HIV, meaning that measuring HIV status is not feasible. This also means that we do not know the starting incidence of HIV and cannot gauge whether a change in HIV infection rates would be detectable. Of course we could make guesses to do a power analysis, but given the cost of the infrastructure we would need for an HIV biomarker study, we want to know the real incidence in our sample before choosing that indicator. We can postpone using HIV status as an indicator until the next round of the evaluation. If the program proves to be effective based on the indicators we do have, it might be worthwhile to do a biomarker follow-up study to determine the program's effects on HIV rates, which is an important consideration for replication and scale-up. Table 5.4 shows a range of possible indicators for this study assessed as to whether they are observable, feasible, and detectable.

TABLE 5.4 Potential indicators and their measurability

		Measurable		
Outcome	Indicators	Observable	Feasible	Detectable
Amount of HIV education	Number of hours teachers spent on HIV curriculum	Yes	Yes	Yes
Modes of HIV education	Use of lecture time to teach about HIV	Yes	Yes	Yes
	Use of playtime to teach about HIV	Yes	Yes	Yes
Knowledge of safe practices	Number of correct answers on test	Yes	Yes	Yes
Unsafe sexual behavior	Number of unprotected sexual encounters	No	No	No
	Incidence of childbearing	Yes	Yes	Yes
Protected sex	Condom use	No	No	No
	Self-reported condom use	Yes	Yes	Yes
HIV infection	HIV status	Yes	No	Unknown

Precision

The more exhaustive and exclusive the indicator, the more precise it is.

Exhaustive indicators An outcome can be measured by more than one indicator, all of them logical ramifications of the outcome. More exhaustive indicators capture more instances of the outcome we need to measure. They improve how accurately we can measure impact. Imagine that a microfinance program increases savings and that it leads to 100 instances of saving. The more comprehensive our indicators, the more of these instances they would capture. We need to think through all the possible manifestations of the concept "savings" in the program's context. People may prefer saving by investing in durable assets rather than by saving money in a bank account. They may buy a goat, buy jewelry, bury money in a jar, or deposit money in a bank. If we measured only money deposits in a bank, we would miss all the other instances of saving. The indicator "money deposited at the bank" would not be an exhaustive indicator of our "savings" outcome.

Is our childbearing indicator an exhaustive indicator of unsafe sexual behavior in our HIV education example? As Figure 5.1 shows, the concept "unsafe sex" has many other effects in the real world besides childbearing. In using childbearing as an indicator of unsafe sex, we are following only the path, marked in boldface, from unsafe sex to vaginal sex to pregnancy to childbirth. Only unprotected vaginal sex leads to pregnancy, and it does so only some of the time. Exhaustiveness concerns how much childbearing captures the incidences of unsafe sex. The question to ask is "If there are 100 instances of unsafe sex in the population, how many result in childbirth?"

In contrast to our childbearing indicator, HIV status is an exhaustive measure of HIV infection. If there were 100 instances of HIV infection and we tested all 100 people at the right time, we would probably have close to 100 HIV-positive tests. HIV status captures all the instances of HIV infection.

Exclusive indicators An exclusive indicator is an indicator that is affected by the outcome of interest and by nothing else. Tears are an example of a nonexclusive indicator. Tears come when people are happy, when they are sad, when they are in pain, and when they cut onions and their eyes become irritated, and some people (actors and politicians) can cry on cue. On their own, tears are an ambiguous sign

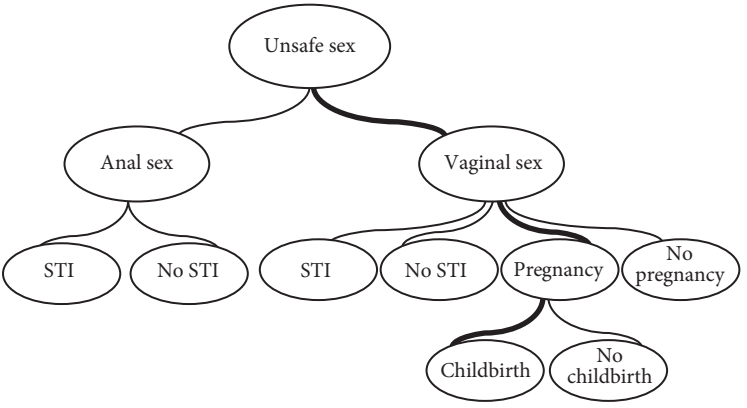

FIGURE 5.1 A logical tree for "unsafe sex"

Note: STI = sexually transmitted infection.

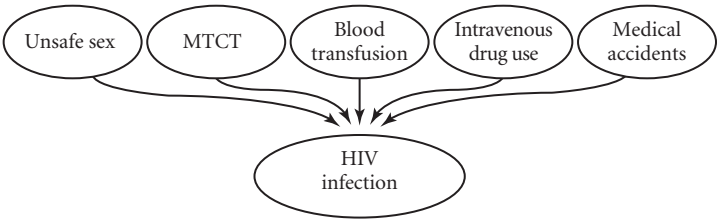

FIGURE 5.2 Possible sources of HIV infection

Note: MTCT = mother-to-child transfer.

of the state of an individual's feelings. If our outcome of interest was sadness, tears would not be a precise indicator.

Although childbearing is not an exhaustive indicator of unsafe sex, it is (in the Kenyan context) exclusive. If we see childbearing, we know for sure that there was an instance of unprotected sexual behavior. Is HIV status an exclusive indicator of unsafe sex? This depends on the context (Figure 5.2). It is possible to become HIV positive without having had unsafe sex. HIV infections can also result from mother-to-child transfers or contact with contaminated blood (through transfusions, intravenous drug use, and medical accidents), although it is rare for 15-year-olds in the Kenyan context to be HIV positive as a result of these alternative transmission routes.[6]

Reliability and social desirability bias

An indicator is reliable when it is hard to forget, counterfeit, or misreport. Say, for example, that we are interested in the incidence of STIs. We could ask people if they have an STI, or we could test them for STIs. People may not tell the truth about their STI status, perhaps because they claim to be abstinent. Test results, on the other hand, are hard to counterfeit. When possible, it is better to choose an indicator that is harder to counterfeit.

Maybe we want to evaluate an HIV prevention program that promotes abstinence and condom use. Neither of these can be directly observed, which leaves us relying on reports from respondents. There are two possible threats: forgetting and deliberate misreporting.

6. This study by Esther Duflo, Pascaline Dupas, and Michael Kremer is summarized as Evaluation 4 in the appendix.

People could simply forget whether they used a condom. To improve the reliability of the variable, we can narrow the question and make it more specific. We could ask, "In how many of your last 10 sexual encounters did you use a condom?" A more reliable question would be, "Think of the last time you had sex. Did you use a condom?"

People could also deliberately misreport. Respondents may lie because they suspect what the "right" answer should be. For example, in contexts in which social stigma is attached to unmarried individuals' being sexually active, some respondents may deliberately lie about being virgins. For example, a study in Kenya found that 12 percent of women who reported being virgins were HIV positive and most had other sexually transmitted infections, making it unlikely that they acquired HIV nonsexually.[7] It may also be the case that people suppress memories that embarrass them or make them unhappy; in other words, they may not even know they are misreporting. When programs encourage certain practices, participants may have a particularly strong desire to report that they are doing these "right" things, so they will report that they are abstaining and using condoms, saving, and not discriminating against their daughters, among other encouraged and socially desirable behaviors. When people misreport because they want to show that their behavior is aligned with what they perceive as socially desirable, the resulting distortion is called *social desirability bias*.

A desire to say what is socially desirable can lead to overreporting as well as underreporting. For example, if microfinance clients are told to invest what they borrowed in a business, clients may state that this is how they used their loans, even if many used them to pay off other loans instead.

Proxy indicators A good way to get around self-reporting bias is to use a proxy indicator that can be measured directly. If the outcome changes, we can predict that the proxy and the unobservable indicator will both change.

For example, the childbearing indicator of unprotected sex used in the HIV education study discussed above is a proxy that helps us avoid social desirability bias. This is particularly important given that

7. J. R. Glynn, M. Caraël, B. Auvert, M. Kahindo, J. Chege, R. Musonda, F. Kaona, and A. Buvé, "Why Do Young Women Have a Much Higher Prevalence of HIV than Young Men? A Study in Kisumu, Kenya and Ndola, Zambia," *AIDS* 15 (2001): S51–60.

the HIV education program being tested promoted abstinence before marriage and fidelity in marriage, so there is a risk that teenagers in the treatment group will be even more likely to misreport unprotected sex than those in the comparison group. Childbearing is a good proxy because it is difficult to counterfeit.

Ensuring that incentives are properly aligned for good reporting We must be careful that we do not use as an outcome something that is incentivized for the treatment group (but not for the comparison group) or for which one group faces a higher incentive than the other. For example, imagine that we are evaluating a program that uses school attendance records to incentivize teachers. Bonuses are given to teachers whose classes have the highest attendance rates as recorded in class registries. Class registries are not a reliable data source in this context. The same problem holds if the incentive is present in both groups but one group faces a higher incentive than the other (for example, if teachers in the treatment group receive 100 rupees for each day they are recorded as present, while comparison group teachers receive 20 rupees for each recorded presence).

As soon as a system is used as the basis of an incentive, we have to worry that there will be manipulation of the data. The problem is even worse if the manipulation is likely to be correlated with our treatment groups. In this case, those who have the incentive (or have a higher incentive) are more likely to manipulate the data. Instead we need to use as our indicator an independent source of information on attendance (such as spot checks). If the independent source shows that the data on which the incentive is based have not been manipulated, we can potentially use them as a supplement to our independent data but we will always need some independent data in these cases.

*Ensuring that data are being collected in an identical manner in the treatment and comparison group*s All the characteristics of what makes a good measure that we have discussed are important regardless of which type of study we are doing. However, for randomized evaluations there is an additional consideration. Everything about the data collection (who does it, how it is done, how frequently, and with what tools) must be identical between the treatment and comparison groups. Often we may have data for the treatment group (from running the actual program) that we do not have for the comparison group. For example, a program in India incentivized teacher attendance at schools

with cameras that produced date- and time-stamped images. The researchers thus had very detailed attendance information for the treatment group. They could not use this same method of cameras to collect attendance data in the comparison group because that mechanism was part of the treatment. Instead the researchers conducted random spot checks to measure teacher attendance. These spot checks were conducted in an identical manner in both treatment and comparison schools.[8]

Field testing

There is only so much progress that can be made toward developing a good data collection plan in the office. Any plan, however well thought out, needs to be tested in the field. It is not enough to do a desktop assessment, because even indicators that make sense and have been used with success elsewhere may still fail in this evaluation, in this context, with this population. This module discusses some considerations for conducting a field test and checking whether our assumptions about our indicators were right.

Have we chosen the right respondents?

Field testing may reveal that the respondents we have chosen do not know the information we are asking about. For example, in an agriculture program it may be reasonable to assume that the head of the household would know about planting decisions. But if in our context men and women are responsible for different crops, the male head of the household will know only about men's crops.

Alternatively, we may find that there are key respondents who know the outcomes of many individuals or of the community as a whole, so we don't have to ask every individual all of the questions but can get the information from a centralized source. The distance to the nearest clinic is a good example of such an indicator.

Do our instruments pick up variation?

The data collection instrument must be sensitive to the variation found in the population where the evaluation is done. If measuring test scores, the test should not be so easy that most participants score 100 percent or so hard that most score 0 percent. Field testing the instrument will help make sure that this is the case.

8. Esther Duflo, Rema Hanna, and Stephen Ryan, "Incentives Work: Getting Teachers to Come to School," *American Economic Review* 102 (2012): 1241–1278.

Is the plan appropriate given the culture and the politics of this context?

If, for example, our program is designed to promote collective action, we will want to ask questions about specific types of collective action that make sense in the specific cultural context. What are the types of activities that cohesive communities in this part of the world engage in collectively? Do they improve local schools? Collectively tend community farm plots? Plan community celebrations? Some of these insights can be gained from talking to experts and reading the literature, but there is no substitute for spending time in the communities, doing qualitative work, and then developing and testing good instruments.[9]

Are the questions phrased in a way that people understand? A question may make sense to us but not to a participant. A common measure of social capital is the number of groups a person belongs to. Early in the field testing of an evaluation in Sierra Leone, we included a question about how many social groups respondents belonged to, but we quickly realized that the question was not well understood. One man claimed to belong to no social groups. After some general discussion it emerged that he was an active member of the local mosque. In addition to attending a prayer group there, he and some of the other elderly members of the mosque had joined together to buy seeds to work some communal land. It became evident that we needed to ask specifically about membership in common groups if we wanted a complete answer to our question about the number of groups someone belonged to.

Are administrative data collected? Do they seem reliable? We may wish to rely on administrative data, such as child enrollment data held by schools or police records. Often it is required by law that these types of records be collected, but that is no guarantee that they exist, and it is certainly no guarantee that they are accurate. Field visits will give us an indication of how well these records are kept.

9. For an example of detailed questions on collective action specific to the local context, see the study by Katherine Casey, Rachel Glennerster, and Edward Miguel, summarized as Evaluation 15 in the appendix. All the survey instruments for this evaluation are available through the J-PAL website at http://www.povertyactionlab.org/evaluation/community-driven-development-sierra-leone.

Is the recall period appropriate? The field test can also help us figure out the right recall period. Is it better to ask about outcomes over the last week or over the last year?

Are the surveys too long? A field test will help us determine how long it takes the average respondent to go through our questions. Fatigue can decrease the quality of the data. But remember that with training and time, our enumerators will get better at administering the questionnaire and the completion time per respondent will fall.

What is the best time and place to find the respondents? The respondents may be less available at some times than at others to answer questions. Field testing will help us plan our survey timetable by answering these questions: How many people will we be able to reach before they go to work? How many can we survey during the middle of the day? Are there days of the week when people are more likely to be found at home? How many times will we have to go back to find the people we missed?

MODULE 5.4 A Catalog of Nonsurvey Instruments

This module gives examples of nonsurvey instruments we can use to quantify difficult-to-measure outcomes such as corruption, empowerment, and discrimination. For each instrument, there are trade-offs to be made between richness of data, cost, and avoidance of reporting bias.

Direct observation

In many cases we can choose to directly observe the behavior of interest in real time, when it happens. We can use random spot checks, mystery clients, or incognito enumerators or can observe group behavior.

All the direct observation instruments discussed here share the advantage of reducing misreporting due to poor memory and deliberate misrepresentation. They can also provide rich, detailed data. However, all direct observation instruments suffer from the drawback of capturing one very specific point in time.

Random spot checks

In a random spot check, an enumerator directly observes and records the indicator of interest at the study site. The timing of the visit is

selected randomly. The idea is to capture the normal state of things. Although any single visit may pick up the unusual behavior of any individual being observed, with a sufficient number of spot checks the average values observed should be a good reflection of reality. For example, a single random spot check is not a good way to assess whether an individual teacher attends regularly. For this we would want to visit at different times on different days. However, if we visit all the teachers in our sample once, visit different teachers at different times on different days of the week, and have a large enough sample of teachers, overall we will get an accurate reflection of absenteeism in our sample. The more frequent the visits, the more expected the visits become and the more polished and less normal the behavior being observed.

When are they useful? Random spot checks are useful when we are measuring outcomes subject to corruption or when participants have incentives to hide their failures or absence. One spot check can also yield rich data on a number of outcomes in one short visit, because we can also observe what an individual is doing, not just whether he is there.

Limitations To ensure that our spot checks give us an accurate picture of reality, we have to conduct a large number of them. These are expensive and require highly trained and well-monitored enumerators. The spot checks also have to measure something that can be quickly observed before people can change their behavior in response to seeing the enumerator.

Example: Spot check of teacher attendance and effort Spot checks have been used to check on the attendance and participation of both providers and beneficiaries. For example, an education project in Indian schools used spot checks to measure teacher and student attendance and teacher effort.[10] Given that the evaluation was sending an enumerator to a school to check whether the teacher was present, she could also quickly record three indicators of teacher effort during an unannounced visit: Were the children in the classroom? Was there anything written on the board? Was the teacher teaching? As for the

10. This study by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden is summarized as Evaluation 2 in the appendix.

students, a roll call was taken on each visit to see who was present. The data from these spot checks are likely to be more reliable than the attendance data recorded in the teacher's register because the teacher may face an incentive to misreport student attendance.

The spot check worked well to check teacher attendance and teaching behavior because these were one-room schools with one teacher. The enumerator can arrive and immediately see what the teacher is doing. In a larger school, a spot check may be able to measure whether teachers are present, but word may spread as the enumerator goes from classroom to classroom, and teachers may quickly start teaching, even if they were talking to colleagues in the staff room when the enumerator arrived.

For further reading

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–1278.

Example: Spot check of environmental audits of industrial pollution An evaluation in Gujarat, India, tested whether changing the incentives of third-party environmental auditors would make them report more accurately. These auditors visit industrial plants three times a year to measure pollution, which they report to the environmental regulator. The evaluation randomly selected some audit visits and had independent agencies spot check them by taking the same pollution readings at the same plant that had just been audited. The person doing the back-check would enter the plant and immediately go to the pollution source (e.g., a boiler stack or chimney) to take a reading before the firm could shut down machinery or take other actions to lessen the pollution it was emitting. These back-checks enabled the researchers to see whether the auditors were telling the truth in their audit reports and formed a valid indicator of misreporting by third-party auditors.

For further reading

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2012. "Truth-Telling by Third-Party Auditors: Evidence from a Randomized Field Experiment in India." Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Mystery clients

The data collection process itself can be made incognito with mystery clients. An enumerator visits the location of interest and pretends to

be a typical client or citizen. The enumerator can then observe and record the quality of the service he receives and any other indicators of interest. In some cases, the "mystery client" can take the form of a paper application sent to an organization instead of a visit by a live person.

If the people the enumerator is visiting and observing do not know that they are part of a research study, the researcher usually has to request a waiver from his or her institution's IRB. A waiver is usually required for not acquiring informed consent before collecting data on someone and for engaging in deception—that is, claiming to need a service when in fact the objective is to collect data. Having the mystery client disclose her purpose after the interaction is complete and not collecting identifying information (i.e., names, locations, or job titles) can help ensure that using mystery clients is compatible with IRB rules. More discussion of IRB rules and judgments as to when research is ethical can be found in Modules 2.4 and 4.2.

When are they useful? Mystery clients are particularly useful for measuring antisocial or illegal activities, such as discrimination or corruption, to which individuals or institutions will not otherwise admit. The use of mystery clients enables us to carefully control an interaction. For example, we can send in mystery clients of different races or genders or have them wear different clothes to gauge how these characteristics affect the interaction.

Limitations If mystery client enumerators must disclose their purpose at the end of the visit, people may change their future behavior in response to knowing that they have been observed and may tell other participants in the study that there are mystery clients in the area.

Example: Using mystery clients to measure the quality of police responsiveness in India An evaluation in Rajasthan, India, tested the effectiveness of different policing reforms on the responsiveness and service of the police. To measure whether police responsiveness to citizens' reports of crime had improved, the researchers sent enumerators on unannounced visits to police stations, where they attempted to register cases posing as victims of various types of crimes. The enumerator did not disclose his identity except when it appeared that the police would actually register the case (to avoid registering a false case,

which is illegal) or when circumstances required that the enumerator disclose his identity—for instance, if the police threatened to prosecute him for filing a false case. The enumerator then recorded his success or failure in registering the crime, the attitudes and actions of the police, and other details, such as the total time taken and the names of the officers with whom he interacted.

This use of mystery clients provided rich indicators of police responsiveness. However, when the police at a station knew that they had been observed in this way, they changed their behavior. In fact, the decoy visits, which had been intended only as a means of collecting outcome data, ended up having a greater effect on the likelihood that a future crime would be registered than did the actual interventions the evaluation tested.

For further reading

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2012. "Can Institutions Be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police." NBER Working Paper 17912. National Bureau of Economic Research, Cambridge, MA.

Example: Measuring race discrimination in hiring processes in the United States Racial, gender, and ethnic discrimination are difficult to measure in the United States because they are illegal. A randomized evaluation sent decoy resumes to firms to see whether employers discriminate based on race in inviting candidates to interview for open positions. The researchers mailed identical resumes to different employers, but some were randomly selected to have a stereotypically white name at the top (e.g., Emily or Greg), while others were randomly selected to receive a resume with a stereotypically African-American name (e.g., Lakisha or Jamal). They then counted the number of callbacks to numbers given for the white-sounding names and the African-American-sounding names. Although discrimination is not directly measurable, these callbacks were, and they formed a good proxy indicator of discrimination.

For further reading

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Incognito enumerators (ride-alongs)

An incognito enumerator experiences the process we wish to measure and records indicators of the outcomes of interest during the process. The enumerator observes the process firsthand and simply counts the number of incidences in a process. Unlike a mystery client, the ride-along does not pose as a client, ask questions, or actively engage in the process she is observing.

When are they useful? These are useful when we need to see a whole process and a single "snapshot" is not enough. Ride-alongs can help us measure the quality of service the participants experience, such as how long it takes to be served or how many steps there are in the process. They are also often used to get an accurate estimate of corruption.

*Limitation*s Ride-alongs can be used only when having a person observing won't change the process. They are expensive because they require extensive training and monitoring, and the process of observation takes a considerable amount of time. We must also come up with an objective way to compare observations across instances and across different enumerators. For example, we can randomly assign enumerators to locations so that differences in how enumerators interpret a script or record events do not bias our data.

Example: Riding along with truckers to record bribes paid at road blocks
A study of the corruption of traffic police in Indonesia used this method. The enumerators, traveling incognito, simply rode along with the truckers and counted the number of solicitations for bribes and the number and amount of bribes paid at each roadblock.

For further reading

Olken, Benjamin, and Patrick Barron. 2009. "The Simple Economics of Extortion: Evidence from Trucking in Aceh." *Journal of Political Economy* 117 (3): 417–452.

Observing group interaction and structured community activities

An enumerator can visit a scheduled event that lets him or her observe group interactions. In most cases, however, there will not be a prescheduled event that occurs in both treatment and comparison communities that we can plan to attend. In this case we need to prompt the

interaction we wish to observe. This approach is known as structured community activities (SCAs).

When is this useful? Sometimes the outcomes that interest us can best be observed when a group of people are interacting. For example, if we are interested in power dynamics or empowerment, observing how people speak or relate to each other can give us information we could not glean from self-reported information in a survey or even from observing individuals.

Limitations This approach works only when the presence of an enumerator is unlikely to change the outcomes we are measuring. Additionally, the type of interactions we want to observe may occur only rarely, so if we simply turn up and wait, we may have to wait a long time before seeing the interaction we want to observe. Unless an interaction, such as a meeting, follows a set schedule, we may have to prompt the interaction that we want to observe.

Example: Prompting a community meeting so we can observe and measure empowerment In cases in which there are no previously planned events to observe, the researchers can prompt an event to occur and use that occasion to observe group interaction. For example, researchers in Sierra Leone conducted SCAs to measure the extent to which a community-driven development program was successful in increasing the participation of women and youth in community decision-making. There were no formally scheduled meetings. Instead a village chief called a meeting whenever a decision had to be made on issues such as how to repair a bridge or how much to contribute to local public works projects.

The researchers therefore created a situation in which the community would have to make a decision. The night before the survey started, the enumerators announced that they would like to address the community as a whole and asked the chief to gather community members the next morning. At that meeting the enumerators said that to thank the community for taking part in the household survey they would like to give the community a gift of either batteries or small packages of iodized salt. They then stood back and let the community decide which gift they preferred. Unobtrusively they observed how many women and youth attended and how many spoke at the meeting. They also noted whether a small group of elders at any point

broke away from the main meeting to "hang heads," that is, deliberate privately about which gift they wanted the community to choose.

For further reading
This study is summarized as Evaluation 15 in the appendix.

Other nonsurvey instruments

When we cannot directly observe the behavior we want to measure, we can use one of many other nonsurvey instruments to record the outcome of interest.

Physical tests

When are these useful? Physical tests are not subject to deliberate misreporting or psychological bias in reporting. A wide variety of physical tests can be used to measure a range of different behaviors and outcomes.

Limitations Physical tests measure one specific outcome. Some tests have high error rates and can produce a noisy (although unbiased) measure. Some physical tests can be difficult to perform in the field and may require specialized technical knowledge. They can also be expensive.

Example: Audit of materials used in public infrastructure projects A program in Indonesia measured corruption on road construction projects. Corruption on construction projects can take the form of diverting funds or using inferior or fewer materials that cost less than the amount allocated and siphoning off the rest of the funds for private use. It can also take the form of diverting the materials themselves, using fewer materials and siphoning the rest for sale or for private use. Here the audit took the form of digging up randomly selected spots on the road and measuring the amount of material used. These data were then used to estimate the total quantity of material used and compare it to official records of expenditures on materials and labor. The gap between the two formed a quantitative indicator of corruption.

For further reading
J-PAL Policy Briefcase. 2012. "Routes to Reduced Corruption." Abdul Latif Jameel Poverty Action Lab, Cambridge MA.
Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–249.

Example: Spot check to test whether chlorine is being used in home drinking water An evaluation in Kenya tested the impact of giving households small bottles of chlorine with which to purify their water. To see whether households were using the chlorine, enumerators tested the drinking water in the family's home to see if residual chlorine was present.

For further reading

Kremer, Michael, Edward Miguel, Sendhil Mullainathan, Clair Null, and Alix Peterson Zwane. 2011. "Social Engineering: Evidence from a Suite of Take-up Experiments in Kenya." Working paper, Harvard University, Cambridge, MA. <http://elsa.berkeley.edu/~emiguel/pdfs/miguel_chlorine dispensers.pdf>.

Biomarkers

Using biomarkers is a highly objective way to test the effects of a program, because biomarkers cannot be manipulated by the subject or respondent. The biomarker used may be directly related to the program (a deworming program checks for worm load) or may be an outcome of a social program that did not directly target health (an education program tests for changes in child health). Some tests of biomarkers include the following:

- STI and HIV tests
- Diagnostic tests for illnesses such as malaria or tuberculosis
- Measures of weight, height, body mass index (BMI), arm circumference
- Tests of toenail clippings for substances such as arsenic
- Pregnancy tests
- Urine tests
- Saliva swabs for cortisol (a stress hormone)

When are these useful? Biomarkers can be much less biased than self-reported information.

Limitations Collecting biomarker data can be very expensive and logistically complicated. There is a risk of high refusal rates if collecting the biomarker is intrusive or painful. We may also be under an ethical obligation to treat those we find have serious medical condi-

tions through biomarker testing. Especially given the expense, it is important to assess in advance whether there will be sufficient variation in the biomarker we choose for us to detect differences between treatment and comparison groups. (See Chapter 6 on statistical power.)

Example: Anemia and a decentralized iron fortification program An evaluation in India tested the impact of a program that sought to address anemia by giving a fortified premix to local millers to add to flour when they ground it for local farmers. Blood samples were taken from respondents to test for anemia. However, many respondents refused to have blood samples collected, and there were higher rates of refusal in comparison communities than in treatment communities (where community members felt thankful for the program and were more willing to cooperate with the enumerators). This high and differential rate of attrition made it hard to interpret the results.

For further reading

Banerjee, Abhijit, Esther Duflo, and Rachel Glennerster. 2011. "Is Decentralized Iron Fortification a Feasible Option to Fight Anemia among the Poorest?" In *Explorations in the Economics of Aging,* ed. David A. Wise, 317–344. Chicago: University of Chicago Press.

Example: BMI as a measure of how families invest in girls Researchers tested the impact of sending job recruiters to villages near Delhi to advertise positions in the business outsourcing industry. When families learned about these well-paying jobs for educated young women, they changed how they invested in their daughters. In addition to educational attainment, girls' BMI increased, which was a good indicator that families invested more in the nutrition and/or health of girls.

For further reading

Jensen, Robert. 2012. "Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India." *Quarterly Journal of Economics* 127 (2): 753–792.

Mechanical tracking devices

Mechanical tracking devices such as cameras or fingerprinting devices can be used to track the attendance of teachers, students, doctors, nurses, or other service providers or to record events or activities. As they have become cheaper and cheaper, mechanical devices have found a lot of use in data collection.

When are these useful? Mechanical devices can overcome problems of distance and timing: a GPS unit can track where someone is on a continual basis without the need for constant observation. For example, a teacher and a principal could easily collude in submitting inaccurate timecards, but an automatic system that relies on a fingerprint to clock the teacher in and out of work prevents corruption.

Limitations Putting a tracking device on someone can alter his behavior, so we need to use tracking devices similarly in the treatment and comparison groups. Moreover, we must understand that the members of our comparison group are not the same as the rest of the population because they know that they are being tracked. Tracking devices can break. If they are intentionally destroyed, breakages may vary by treatment and comparison group.

Example: Measuring road quality with video cameras on cars in Sierra Leone In Sierra Leone researchers studied the impact of a road-building project. To get an objective indicator of road quality and to measure road usage, they mounted video cameras with GPS units on cars. They used the cameras to measure the speed of the car, the width of the road, and how much traffic was on the road.[11]

Spatial demography

GPS readings, satellite images, and other artifacts from spatial geography can be useful in data collection. For example, GPS devices can be used to accurately measure the distance between places (e.g., between a house and its nearest school) or to measure the size of a farm.

When is this useful? Spatial data allow us to make more use of distance in our analysis. For example, we can use distance to predict take-up rates. We can also use spatial data to measure spillovers. For example, we can define villages within a fixed radius of a treatment village as spillover villages and those farther from the treatment villages as pure control villages. Satellite imagery allows us to collect data over much larger areas than would be feasible using survey data.

11. Lorenzo Casaburi, Rachel Glennerster, and Tavneet Suri, "Rural Roads and Intermediated Trade: Regression Discontinuity Evidence from Sierra Leone," 2012, http://ssrn.com/abstract=2161643 or http://dx.doi.org/10.2139/ssrn.2161643.

Limitations When we use distance in our analysis, we are often interested in the travel time between two points. Distance as measured by GPS does not always match travel time—for example, if roads, rivers, or difficult topography connect or separate two locations. GPS information is not always 100 percent accurate because of bad weather during data collection, poor reading of the GPS output by enumerators, or poor-quality machines. Using a GPS device that digitally records the reading (rather than having enumerators write it down) is recommended, but if this is not feasible, enumerators must be thoroughly trained to use the GPS devices appropriately. Satellite imagery is of very mixed quality in different parts of the world, detailed in some (mainly rich countries) and limited in many poor countries. The availability and quality of satellite imagery are improving over time.

Example: Satellite imagery to measure deforestation Research in Indonesia used data from satellite imagery to track changes in deforestation. Researchers constructed a data set that combined satellite imagery with GIS data on district boundaries and land-use classifications. This enabled them to construct a data set that captures deforestation across localities.

For further reading

Burgess, Robin, Matthew Hansen, Benjamin Olken, Peter Potapov, and Stefanie Sieber. 2011. "The Political Economy of Deforestation in the Tropics." NBER Working Paper 17417. National Bureau of Economic Research, Cambridge, MA.

Example: Using GPS to estimate the effect of distance on program take-up An evaluation in Malawi tested how people respond to incentives to learn their HIV status and the impact of distance on take-up of learning one's HIV status. The researcher used GPS data on the location of the study participants in order to group households into zones. Then the location of the temporary HIV test results center was chosen randomly within each zone. This allowed the researcher to see how distance to the center might reduce the probability that people would collect their test results.

For further reading

J-PAL Policy Briefcase. 2011. "Know Your Status?" Abdul Latif Jameel Poverty Action Lab, Cambridge, MA.

Thornton, Rebecca L. 2008. "The Demand for, and Impact of, Learning HIV Status." *American Economic Review* 98 (5): 1829–1863.

Participatory resource appraisals

These appraisals use a number of dynamic approaches to incorporate local knowledge and opinions in the data collection process. They are used, for example, to find out the amount of resources that are available in a village, or to estimate the wealth of a given family, or to decide who is eligible to be a program beneficiary.

When are these useful? A participatory approach is useful when interaction between people can give us more information than asking people questions individually. Each individual person may only know part of the picture, but as they talk they prompt each other, and something one person says reminds another person of more details. This can be useful when we want information from a long time period (e.g., the village's investments in public goods over the past five years), and asking one person at a time may produce less accurate data than asking a group at once.

Limitations One concern with participatory appraisals is that people may not be willing to make certain comments in front of others. In particular, if it is likely that members of the community elite will be present in the group, it may be difficult for nonelite members to criticize them openly. Participants may not be willing to contradict a neighbor or admit that they don't know the answer to a question. Choosing the right group of knowledgeable participants and asking the right questions is important for participatory appraisals to work well.

Example: Recording when and where wells were built and how they were maintained Researchers wanted to test whether quotas for women in local politics affected how public goods were supplied. They wanted to see whether there were more wells or whether existing wells were better or worse maintained in villages where the chief village councilor position had been reserved for a woman. In each village, no single person would know the history of all the wells in the area. However, enumerators could gather a group of informed people from the village and engage them in a conversation about the wells. They drew up maps of the village, and as people shared information on where and when wells were built and when they were repaired, they added

that information to the maps. One person might say that the well near the school was dug five years ago, which prompted another to mention that it was actually six years ago, because it was dug before her child left school. This reminds a third person to mention that the well was repaired the previous spring. At the end of the conversation, the enumerators could select some of the villagers for a walkabout to check the accuracy of the recollections of the resources on the map and note the condition of the wells. The key is that interaction between the local people as they prompted and corrected each other helped a much more accurate and comprehensive picture to emerge.

For further reading

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72 (5): 1409–1443.

J-PAL Policy Briefcase. 2006. "Women as Policymakers." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/women-policy-makers.

This study is also summarized as Evaluation 12 in the appendix.

Example: Gathering tracking information on program participants An HIV education program in Kenya targeted adolescents in the final years of primary school. The main outcome was a biological proxy for unsafe sexual behavior: the incidence of pregnancy and childbearing among girls. The data were collected during six visits to the schools in the three years following the intervention.

To limit attrition in this extended follow-up, the researchers collected tracking information on program participants by conducting participatory group surveys of students enrolled at each participant's former school. At each visit the list of all participants in the original baseline sample was read aloud to students enrolled in the upper grades. For each participant, a series of questions was asked to prompt conversation about the student: Is she still going to school? If yes, to which [secondary] school, in what grade? Does she still live in the area? Is she married? Is she pregnant? Does she have any children? If yes, how many children does she have? One person might remember that the girl moved to a different area. This reminded another person that when he last saw her, she was pregnant and no longer attending school.

The participatory process connects bits of information that "live" with different people. In this case, the participatory group survey gen-

erated very accurate data. In a subsample of 282 teenage girls who were tracked at their homes and interviewed, 88 percent of those who had been reported to have started childbearing had indeed started, and 92 percent of those who had been reported as not having started had not started. The accuracy rates were similar across treatment groups.

For further reading

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2012. "Education, HIV and Early Fertility: Experimental Evidence from Kenya." Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Dupas, Pascaline. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3: 1–34.

J-PAL Policy Briefcase. 2007. "Cheap and Effective Ways to Change Adolescents' Sexual Behavior." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA.

This study is also summarized as Evaluation 4 in the appendix.

Using purchase decisions to reveal preferences

Sometimes difficult-to-observe behavior or preferences can be inferred from how individuals choose to spend money. Although someone may give an inaccurate answer on a survey to please the enumerator, it is likely that when people's own resources are on the line their purchasing decisions will better reflect their true preferences. People may still buy a token amount to please the enumerator, but overall it is likely that when people spend their own money, their purchases reflect their true preferences more closely than does self-reported information in a survey.

When is this useful? Using purchase decisions is useful when we think that there is a strong social desirability bias in people's answers to survey questions about how much they value a good or how often they would buy it. For example, if we ask in a survey how often children in a household get sick from unclean water and then ask people how much they value clean water, they are very likely to overestimate the true value they place on it. Similarly, people may be unwilling to admit that they don't use condoms. Observing purchasing decisions allows us to test for demand at prices that may not exist in the market. If we ask about existing purchases, we find out only about demand and the current price. However, we can manipulate the price to see what people would buy at different prices.

Limitations Although spending real money reduces the risk of social desirability bias, it does not eliminate it. People may well buy a token amount to please the enumerator. It is useful, when possible, to offer larger amounts to distinguish between token purchases and substantial purchase.

Example: Inferring sexual behavior from condom-purchasing behavior A program in Malawi provided free HIV tests in a door-to-door campaign and offered small cash incentives to people to collect their results at temporary mobile voluntary care and testing centers in their community. Later interviewers visited the homes of participants and offered them the chance to buy subsidized condoms. Whether people chose to buy the subsidized condoms revealed the impact that learning their HIV status could have on this proxy for sexual behavior.

For further reading

J-PAL Policy Briefcase. 2011. "Know Your Status?" Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/know-your-status.

Thornton, Rebecca L. 2008. "The Demand for, and Impact of, Learning HIV Status." *American Economic Review* 98 (5): 1829–1863.

Example: Learning whether free distribution reduces future purchases of bed nets An evaluation in Kenya tested whether people who had received free bed nets in the past would feel "entitled" to free bed nets indefinitely and thus were less likely to purchase bed nets in the future. In a survey the enumerators could have asked people whether they would purchase a bed net in the future, but the researchers obtained far more reliable data by offering bed nets for sale both to people who had been randomly selected to receive free bed nets in the past and to those who had had to pay for bed nets in the past.

For further reading

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Evaluation Experiment." *Quarterly Journal of Economics* 125 (1): 1–45.

Dupas, Pascaline. 2009. "What Matters (and What Does Not) in Households' Decision to Invest in Malaria Prevention?" *American Economic Review* 99 (2): 224–230.

———. 2012. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." NBER Working Paper 16298. National Bureau of Economic Research, Cambridge, MA.

J-PAL Policy Bulletin. 2011. "The Price Is Wrong." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/the-price-is-wrong
This study is also summarized as Evaluation 6 in the appendix.

Example: Learning about collective action by offering matching funds for community fundraising In an evaluation in Sierra Leone, communities were given vouchers that could be redeemed at a local building supply store if the community also raised its own funding. For approximately every US$2 raised by the community, the voucher would allow the community to receive an additional US$1. Because it was always worthwhile for the community to redeem the vouchers, the number of vouchers redeemed at the store by each community was used as a measure of how well the community was able to coordinate to raise funding for local projects.

For further reading
This study is summarized as Evaluation 15 in the appendix.

Games

Sometimes having the subject play a "game" can measure social qualities such as altruism, trust, and fairness. Normally this is done in a controlled laboratory experiment, with participants' behavior in the game as the outcome of interest. Examples of standard games include the dictator game, which measures altruism, the trust game, and the ultimatum game, which measures fairness or reciprocation.[12]

When are these useful? Games can be useful when we want to test theories of how people will respond to different incentives and scenarios. We can run lots of different scenarios in one day with games. If we tried to run the same range of scenarios in real life, it would take us many years and hundreds of thousands of dollars to run. Games can help us differentiate between different types of people. For example, games can identify people who are risk averse and those who are risk loving.

12. A description of a trust game and an example of how it can be used in a randomized evaluation can be found in Dean Karlan's "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions," *American Economic Review* 95 (December 2005): 1688–1699. A description of the ultimatum and dictator games can be found in Robert Forsythe, Joel Horowitz, N. E. Savin, and Martin Sefton's "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior* 6 (1994): 347–369.

Limitations Unlike most aspects of randomized evaluations, games are not played in the "real world," and thus the findings from them may not be easily generalized to other settings. Participants in games sometimes think they are being judged and play a certain way; for example, they may demonstrate substantial altruism in the hope that this will make them eligible to receive rewards or participate in projects.

Example: Using games to identify attitudes toward risk and impulsiveness An ongoing evaluation in Uganda is measuring the impact of a small grants program on the livelihoods and empowerment of at-risk women. The evaluation will use games to identify specific attitudes and traits in participants to see if the program has a differential impact on people based on these traits. The researchers are trying to assess the role that attitudes toward risk, impulsiveness, altruism, trust, and public mindedness can play in group and poverty dynamics. They are using behavioral games to measure attitudes toward risk, time preferences, in-group trust, honesty, in-group altruism, group coordination, and in-group public goods contributions both before the intervention begins and at the conclusion of the program. The games have three goals: (1) to associate game behavior with individual characteristics (e.g., are wealthier individuals more risk averse?), (2) to correlate game behavior with long-term outcomes (e.g., do more patient individuals benefit the most from business skills?), and (3) to serve as outcome measures themselves (e.g., does group formation increase the level of trust and cooperation?).

For further reading

Annan, Jeannie, Chris Blattman, Eric Green, and Julian Jamison. "Uganda: Enterprises for ultra-poor women after war." Accessed January 2, 2012. http://chrisblattman.com/projects/wings/.

Example: Public goods games to test collective action Researchers conducted a randomized evaluation to test the impact of introducing local democratic governance institutions in northern Liberia. The intervention sought to strengthen the ability of communities to solve collective action problems. Five months after the intervention was completed, the researchers used a communitywide public goods game as a behavioral measure of collective action capacity. They found that treated communities contributed significantly more in the public goods game than did comparison communities. The researchers then used

surveys of the game players to try to understand the mechanisms by which the program had affected contributions.

For further reading

Fearon, James, Macartan Humphreys, and Jeremy M. Weinstein. 2009. "Development Assistance, Institution Building, and Social Cohesion after Civil War: Evidence from a Field Experiment in Liberia." CGD Working Paper 194. Center for Global Development, Washington, DC.

———. 2011. "Democratic Institutions and Collective Action Capacity: Results from a Field Experiment in Post-Conflict Liberia." Working paper, Stanford University, Stanford, CA.

Standardized tests

Giving participants a standardized test is a simple way to measure what the participants learned as a result of the program. The most common example is subject tests given to schoolchildren. Tests do not need to be a pen-and-paper test. Sometimes a paper test will measure skills other than the ones that interest us. (For example, we may be interested in measuring problem solving but end up simply measuring literacy.) Visuals, manipulable puzzles, and other devices can enable testing in which language and literacy do not get in the way of our measurement.

When are these useful? These tests are useful when we want to test a large number of people quickly and we are interested in their knowledge of a specific set of topics.

Limitations Test scores are only an intermediate outcome. What we really want to know is whether a program has led to learning that will improve the lives of children later in life. A particular concern is whether the program makes the relationship between test scores and later life outcomes less reliable. For example, if we think that the program makes it more likely that teachers will teach to the test, test scores become a less reliable outcome measure because there is a systematic difference in how well test scores measure learning between the treatment and comparison groups. We may need to see if results persist or administer other tests that are not incentivized to try to measure reliable outcomes in this instance.

Example: Testing basic math and reading skills in India In India, the NGO Pratham administers standardized tests to assess very basic lit-

eracy and numeracy skills in children. An evaluation of Pratham's remedial education program used simple tests that were targeted to the actual learning levels of the children to measure whether the program improved learning. The set of tests were designed to pick up different levels of very basic skills, ranging from deciphering letters to reading words to reading a sentence. Had the evaluation used tests designed to measure whether children were meeting grade-level standards (i.e., a third-grade test used for a third-grade child), most of the children would have failed the test. From the standpoint of the evaluation, this would not have been useful for detecting changes in learning levels because the tests used should not be so difficult (or so easy) that most children have similar scores (either acing or failing the tests).

For further reading

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–1264.

J-PAL Policy Briefcase. 2006. "Making Schools Work for Marginalized Children: Evidence from an Inexpensive and Effective Program in India." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/making-schools-work-marginalized-children. This study is also summarized as Evaluation 2 in the appendix.

Using vignettes to assess knowledge

Vignettes are hypothetical scenarios presented to respondents to measure their skills.

When is this useful? Vignettes can give more information than a standardized test by seeing how a respondent analyzes a "real-life" scenario.

Limitations Vignettes do not always tell us what people will do when faced with the same situation in real life; when we use a vignette, they know they are being tested.

Example: Assessing doctors' ability to diagnose common illnesses Clinical vignettes were used to measure the knowledge and competence of doctors in India. Doctors were presented with patients' symptoms and given the opportunity to ask questions about their case histories. The cases included examples such as those of an infant with diarrhea, a pregnant woman with preeclampsia, a man with tuberculosis, and a

girl with depression. The doctors were then asked which tests they would prescribe and which treatments they would give. A competence index was constructed based on the questions asked by the doctors and their treatment decisions. Although the test measured how much doctors knew, the authors found that many doctors did better in vignettes than when they faced real patients. The vignettes were useful in pointing out this difference between knowledge and action.

For further reading

Das, Jishnu, and Jeffrey Hammer. 2007. "Location, Location, Location: Residence, Wealth, and the Quality of Medical Care in Delhi, India." *Health Affairs* 26 (3): w338–351.

———. 2007. "Money for Nothing: The Dire Straits of Medical Practice in India." *Journal of Development Economics* 83 (1): 1–36.

Using vignettes and hypothetical scenarios to assess unstated biases

We can create multiple identical vignettes that differ only by the sex or race of the person in the hypothetical scenarios. We can then randomly select which respondents hear which version. If the respondents, on average, rate the vignettes differently based on the race or sex of the person in the scenarios, this is a measure of bias.

When is this useful? Vignettes are useful for measuring unstated or socially undesirable biases and discrimination. They allow the researcher to control exactly what the respondents hear—enabling them to isolate the one difference they want to study.

Limitations This type of vignette is designed to measure the differential response to one very specific attribute, such as gender. This can be an expensive way of measuring one outcome. It is thus most useful when the one outcome is the main focus of the study and when a survey response is likely to be unreliable.

Example: Recorded speeches by male and female leaders in India An evaluation in India measured the effect of quotas for women in local politics. The researchers wanted to measure whether villagers would rate a speech given by a woman lower than an identical speech given by a man. Each respondent heard a short tape-recorded speech in which a leader responded to a villager complaint by requesting that

villagers contribute money and effort. Some respondents were randomly selected to hear the speech read by a male voice and others to hear it with a female voice. They were then asked to evaluate the leader's performance and effectiveness.

For further reading

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124 (4): 1497–1539.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72 (5): 1409–1443.

J-PAL Policy Briefcase. 2006. "Women as Policymakers." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/women-policy-makers.

———. 2012. "Raising Female Leaders." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/raising-female-leaders.

This study is also summarized as Evaluation 12 in the appendix.

Example: Grading bias by caste in India An evaluation in India tested whether teachers discriminate based on sex or caste when grading exams. The researchers ran an exam competition through which children competed for a large financial prize and recruited teachers to grade the exams. The researchers then randomly assigned child "characteristics" (age, gender, and caste) to the cover sheets of the exams to ensure that there was no systematic relationship between the characteristics observed by the teachers and the quality of the exams. They found that teachers gave exams that were identified as being submitted by lower-caste students scores that were about 0.03–0.09 standard deviations lower than they gave exams that were identified as being submitted by high-caste students.

For further reading

Hanna, Rema, and Leigh Linden. 2009. "Measuring Discrimination in Education." NBER Working Paper 15057. National Bureau of Economic Research, Cambridge, MA.

Implicit association tests (IATs)

An IAT is an experimental method that relies on the idea that respondents who more quickly pair two concepts in a rapid categorization task associate those concepts more strongly.

When are these useful? IATs can measure biases that participants may not want to explicitly state or biases that participants do not even know they hold.

Limitations Although IATs allow researchers to quantify biases, they still provide only approximations. They rely on the assumption that immediate categorization results from a strong association between two concepts and that the strong association results from biases or stereotypes. For example, people may be aware that society reinforces stereotypes and may deliberately try to counter their own stereotypes when they face situations in real life.

Example: IATs to measure biases about female leaders in India An evaluation in India measured whether villagers who had been exposed to female leaders through a quota system had increased or decreased their bias against female leaders. During the IAT, respondents were shown sets of two pictures on the left-hand side of a computer screen and two pictures on the right-hand side. They were shown one of two picture configurations. The "stereotypical" configuration showed a picture of a man next to a picture of a leadership setting (such as a parliament building) on one side and a picture of a woman next to a picture of a domestic setting on the other. In the "nonstereotypical" configuration, the picture of the woman was placed next to the leadership setting and the picture of the man next to the domestic setting. Respondents were then played recordings or shown pictures in the middle of the screen that represented different concepts or activities. These could include things like cooking, washing, mechanics, politics, children, and money. Each respondent would then press a button to indicate whether the concept or activity belonged to the right- or left-hand side of the screen. If a participant more strongly associated leadership with men and domestic skills with women, she would more quickly sort leadership activities to the correct side if that side also had a picture of a man and domestic activities to the correct side if that side also had the picture of the woman. The researchers then calculated the difference in average response times between the stereotypical and nonstereotypical picture configurations as the quantitative measure of bias: the larger the difference in sorting time, the stronger the implicit stereotype.

For further reading

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124 (4): 1497–1539.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72 (5): 1409–1443.

J-PAL Policy Briefcase. 2006. "Women as Policymakers." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. http://www.povertyactionlab.org/publication/women-policy-makers.

———. 2012. "Raising Female Leaders." Abdul Latif Jameel Poverty Action Lab,Cambridge,MA.http://www.povertyactionlab.org/publication/raising-female-leaders.

This study is also summarized as Evaluation 12 in the appendix.

Using list randomization to quantify hidden undesirable behavior

List randomization enables respondents to report on sensitive behavior without allowing the researcher to identify individual responses. Half of the survey respondents are randomly selected to receive a short list of activities; they are asked *how many* of the activities on the list they have engaged in, but they are not asked to report which specific activities. The other half of the survey respondents sees the same list of activities but the key sensitive activity (the one of interest to the researchers) is added. Again, respondents report the number of activities on the list that they have engaged in before. The researchers can then average the number of activities respondents in the two groups have previously engaged in. The difference in the average number of activities of the two groups lets the researchers estimate the proportion of respondents who engage in the sensitive behavior.

When is this useful? List randomization can allow us to estimate how common a sensitive, forbidden, or socially unacceptable activity is. This is especially useful when program beneficiaries may fear that telling the truth may disqualify them from receiving program benefits in the future.

Limitations This technique can supply only aggregate estimates of how common an activity is, not individual-level information.

Example: Estimating how many microfinance clients use business loans for personal consumption Researchers conducting evaluations in the

Philippines and Peru used list randomization to estimate how many microfinance clients used loans that were intended for business investment to pay off other loans or to purchase personal household items. For example, one group was presented with a list of different uses for a loan (such as purchasing merchandise or equipment for a business), and the second group received the same list, plus one more statement such as "I used at least a quarter of my loan on *household items,* such as food, a TV, a radio, et cetera." The difference in the average number of activities people reported engaging in between the two groups let researchers estimate the percentage of people who spent their loans on household items, which was much higher than the percentage of people who self-reported that behavior in surveys.

For further reading

Karlan, Dean, and Jonathan Zinman. 2010. "Expanding Microenterprise Credit Access: Using Randomized Supply Decisions to Estimate the Impacts in Manila." CGD working paper, Center for Global Development, Washington, DC.

———. 2011. "List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds." Working paper, Yale University, New Haven, CT.

This study is also summarized as Evaluation 13 in the appendix.

Using data patterns to check for cheating

When people face an incentive or penalty based on data they submit, they may be likely to cheat and falsify the data. Researchers can conduct checks on the data sets to look for certain patterns that indicate cheating. For example, a teacher seeking to boost the test scores of her students may fill in the correct answers to any questions the students left blank when they ran out of time at the end of a section. In the data we would then see a pattern of all correct answers at the ends of test sections.

When is this useful? It is useful to check for cheating by program participants or by the evaluation's enumerators.

Limitations Sometimes patterns may look like cheating and not be. For example, if a class is doing badly on a test and then everyone gets one question at the end right, it may be because of cheating or because the class was taught about that subject very recently and they all remember it. Usually there is no reason to think that these coinci-

dences will be more common in treatment than control groups (or vice versa), but we need to watch out for these possibilities.

Example: Corruption in smog tests in Mexico A study in Mexico tested for corruption in emission checks of cars by running a statistical test for specific patterns in the data of readings from emission checks. In this context, cheating can occur when a customer bribes an emission testing technician to use a clean testing car that has passed the emission test, commonly called a "donor car," to provide the emission readings for the bribing costumer's dirty car. A donor car is needed because emissions cannot be entered manually into the center's computer. The car's information, on the other hand, has to be entered manually into the computer, which allows the technicians to enter the information from a dirty car and match that with emissions from a clean car. When technicians run an emission test multiple times in a row on the same clean car but falsely label those readings as being from dirty cars, the data will reveal a pattern. Consecutive emission readings look similar, despite being labeled as belonging to cars of different years, makes, and models. This can be detected with statistical tests for serial correlation.

For further reading

Oliva, Paulina. 2012. "Environmental Regulations and Corruption: Automobile Emissions in Mexico City." Working paper, University of California, Santa Barbara.

Measuring social interaction and network effects

Collecting rich data on social networks opens a whole set of possibilities for data analysis and research on issues such as peer effects, spillovers, and social diffusion of knowledge. This type of data is typically collected with surveys or participatory appraisals, and it is very important to acquire comprehensive data.

When is this useful? Randomized evaluations are increasingly looking at the question of how technology and information spread within and between communities. Is it possible to train a few people in a community on a new technology or on improved health prevention approaches and have them spread the word to others? How do we identify the relevant people in a community to train or pass information to? In order to answer these questions, we need to collect information on social connections and use this to draw a map of the social

networks in the community. From the basic information on whom different individuals know and talk to—about, say, agriculture—we can describe the social network in a given community. Is a community best described as composed of different groups who share information within their respective group but don't share across groups very much? Or are social links much more fluid, with many connections across subgroups? If there are distinct subgroups, who are the connectors: the people who can pass information from one subgroup to another in a community? If we see that a technology initially introduced to one farmer has spread to another farmer, how close was the new adopter to the original adopter? Within the first two years of the introduction of a new technology, did it spread mainly to farmers who were one or two social links away from the original adopter?

Limitations Although it is relatively straightforward to collect basic information about whether someone has lots of friends, it is hard to perform a full-scale social network analysis. If we want to be able to ask how many social links a technology passed through in a given amount of time, we have to have a complete mapping of every social link in the entire community. In other words, we have to ask every single person in the community whom they know and talk to, and we also have to be able to match the information given by one person about her social contacts with the information given by those social contacts. This matching can be difficult and imprecise when a single person may have different names (for example, a formal name and a nickname), when names are misspelled by enumerators, and when many people in the same village have the same name (as is common in many countries).

A further limitation is that even small errors in data collection can sharply undermine the quality of the network description. Usually when we collect data, small errors creep in: people cannot remember exactly how long ago they were sick or how much rice they planted. Errors also arise when data are entered. But these errors usually even out because some people overestimate and some underestimate, with the result that the errors don't affect averages very much. But because everything is connected to everything else in a network, a small data error (such as confusing Mohammed Kanu with Mommed Kanu) can radically change the connection map.

Imagine a village that is socially very divided—by religion, caste, income, or simply by geography. The true social network is described

below (Figure 5.3A), with each farmer represented by a dot and a lowercase letter and the lines representing people they talk to about agriculture. For information to pass from a to c it has to go through two connections, from a to b and then from b to c. Now imagine that in collecting our data we made one mistake. Farmer a told us that she talks to b, e, and f about agriculture, but we ended up recording that she talks to b, e, and g about agriculture. Our mistaken view of the network is that described in Figure 5.3B. Although in reality information has to pass through a few central figures, our view of the network dramatically shortens the information links between the two groups. And suddenly farmer b has become an important nodal figure through whom information has to pass, when in fact he does not play this role.

Example: Social networks in Karnataka, India Researchers collected data on social networks in 75 villages of Karnataka, India. A census of households was conducted, and a subset of individuals was asked detailed questions about the relationships they had with others in the village. This information was used to create network graphs for each

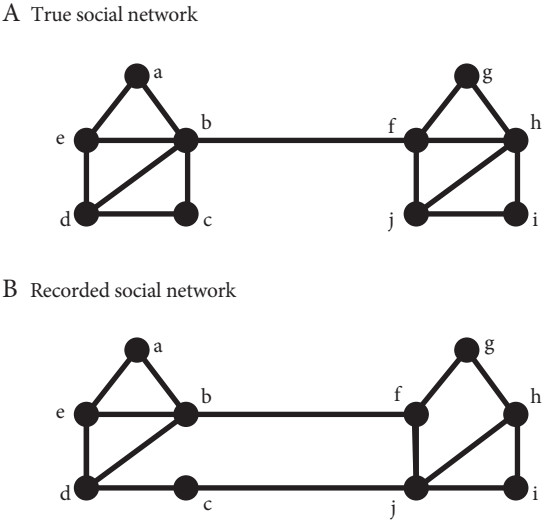

FIGURE 5.3 Social networks

Note: Each dot with a lowercase letter represents a farmer. The lines represent their interactions.

village that could be used to study the diffusion of microfinance throughout the villages.[13] More information about this process, including the original data and a set of research papers that have made us aware of the social networks data, can be found at the MIT Department of Economics website: http://economics.mit.edu/faculty/eduflo/social.

13. Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson, "The Diffusion of Microfinance." NBER Research Paper w17743, National Bureau of Economic Research, Cambridge, MA, 2012.