# Annual Monitoring of Health Outcome Indicators

## Description and Assessment of Alternative Data Collection Methodologies

Prepared by MEASURE Evaluation

for

USAID Global Health Bureau

September 2006

This paper was drafted by Karen Foreit and Stirling Cummings,
with contributions from Jennifer Luna, Anwer Aqil, Philip Setel, Thomas McDevitt,
Martin Vaessen, Gustavo Angeles, Charles Teller, Ellen Starbird, and Alfred Bartlett.

*MEASURE Evaluation is funded by USAID through Cooperative Agreement GPO-A-00-03-00003-00 and is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, in partnership with Constella Futures, John Snow, Inc., ORC Macro, and Tulane University.*

Annual Monitoring of Health Outcome Indicators
Description and Assessment of Alternative Data Collection Methodologies

# Table of Contents

## Introduction

USAID has a long history of successful investment in health-related interventions, based on identifying low-cost, feasible interventions that are scientifically proven to improve health and fertility outcomes in low income countries, and supporting their implementation at scale. Examples include insecticide treated bednets (ITNs); appropriate treatment for malaria, pneumonia, and other diseases that kill children; immunization; modern medical and surgical approaches to contraception; and prevention and treatment of the major causes of mortality in pregnancy and childbirth.

As part of the restructuring of United States Foreign Assistance, the Global Health (GH) Bureau has been asked to provide performance indicators that can be collected annually by country Missions, which can measure USAID's specific contributions within the country and against which Missions can set annual targets. These indicators and targets will be incorporated into the annual country operational plans (OP). For FY 07 and the 35 "fast track" countries, GH has provided output indicators for family planning (FP), maternal and child health (MCH) and Malaria. For Avian Influenza and TB, indicators previously discussed and agreed upon by the US government (USG) partners will be used as is or modified to try to capture only the USG contribution. For HIV/AIDS the PEPFAR indicators will continue to be used.

While recognizing the importance of monitoring program activities, GH strongly argues that program effectiveness is measured by the degree to which these interventions reach people – that is, the degree to which outcomes such as "*children sleeping under ITNs*" are achieved. This strategy has been validated by international scientific studies including a recent comprehensive review of global child health in *The Lancet*[1] and the global Disease Control Priorities Project[2]. As a result, other donors and organizations are refocusing their investments on achieving these outcomes. The US government is seen as a global leader in this movement.

Failure to monitor outcomes on a regular basis can hide fundamental program inadequacies, delay action to improve effectiveness and ultimately lead to wasted investment. For example, monitoring the number of bednets distributed does not capture whether young children and pregnant women are sleeping under those nets; thus while the output (bednets distributed) is necessary for program success, it is not sufficient for the intended health impact (which can be achieved only if the bednets are actually used). This need to monitor at the outcome level is especially true of interventions carried out at the household and community (versus health facility) level, such as use of fluid therapy for dehydrated children and promotion of breastfeeding.

Moreover, there is ample evidence demonstrating that even the best output indicators often correlate poorly with outcomes.

- Immunization estimates based on health service statistics or "doses administered" often differ widely from actual coverage as determined by population-based surveys;
- Contraceptive distribution and its ancillary indicator of "couple-years of protection" (CYP) show at best weak correlations with actual contraceptive prevalence;
- Improving health system variables such as workers trained, supervisory visits made and adequate supplies at the facility level has failed to significantly increase coverage of essential child health interventions in many countries participating in the global WHO "Integrated Management of Childhood Illness" program.

[1] [Lancet reference to come]

[2] Disease Control Priorities Project (DCPP) is funded principally through a grant from the Bill & Melinda Gates Foundation and implemented by The World Bank, the Fogarty International Center of the National Institutes of Health, the World Health Organization, the Population Reference Bureau and the National Library of Medicine.

Therefore, beginning in FY08, GH proposes USAID complement annual monitoring of *output* indicators with periodic (annual or biennial) monitoring of selected outcome indicators for FP, MCH and Malaria (Attachment 1). To this end, GH wishes to determine which tools/methodologies would allow outcome indicators for FP, MCH and Malaria to be collected on a frequent basis, would be relatively inexpensive to implement and would provide regular information at the program level that Missions can use to show progress, set targets and make mid-course corrections.

Objective: Produce simple and validated health sector outcome indicators to monitor USAID supported programs annually (individual countries might choose to collect data biennially).

Strategy: Identify and field test rapid data collection methods and tools to produce outcome indicators for use in annual program monitoring, target setting and on-going evaluation of USAID programs.

The new directives for annual reporting of health indicators parallel those established several years ago under the President's Emergency Plan for AIDS Relief (PEPFAR). Thus, the experience gained under PEPFAR may be instructive. Major challenges under PEPFAR have been the need to get results reported by strictly enforced deadlines and the implications of decentralized approaches for both meeting those deadlines and for ensuring some level of consistency and minimal quality across different groups. Many focus countries found it necessary to invest in a large amount of capacity building of partners to strengthen their M&E skills and their reporting and required external assistance to establish systems to coordinate reporting for the semi-annual and annual report to meet the deadlines. It may be the case that the outcomes indicators could be collected in a centralized way that provides more quality control, but that benefit needs to be balanced against the desire for sustainability and involvement of partners to strengthen local utility for program management.

This background paper is designed to inform discussions among GH, MEASURE Evaluation, and key experts in order to select one or more rapid data collection methods for field testing in selected countries.

## Monitoring and Evaluation systems, indicators and measurement

GH has proposed a list of outcome indicators for Malaria, MCH and family planning. As can be seen in Table 1, all of these indicators are based on binomial variables (each case in the numerator takes a value of either yes or no), but the denominators vary by indicator from household to individuals within the household with specified characteristics.

**Table 1. Proposed Outcome indicators for Malaria, MCH and FP**

| | Indicator | Denominator |
|---|---|---|
| Malaria | Households with at least one ITN | Household |
| | Percent of children under five who slept under an ITN the previous night | Children under five |
| | Percent of pregnant women who received two or more doses of SP for IPT during their pregnancy | Births during reference period |
| | Percent of households targeted for indoor residual spraying that have been sprayed | Household targeted for spraying |
| | Percent of children under five with suspected malaria who received treatment with an antimalarial drug within 24 hours of onset of their symptoms | Children under five with fever during reference period |
| | Percent of vulnerable populations protected by insecticide-treated mosquito nets or indoor residual spraying | Pregnant women and children under five |
| MCH | Percent of women making greater than or equal to 4 antenatal care visits | Live births during reference period |
| | Percent of deliveries attended by skilled attendants | Live births during reference period |
| | Percent of children underweight for age | Children under five |
| | Immunization coverage (DPT3) | Children 12-23 months |
| | Appropriate care seeking for pneumonia | Children under five with symptoms of pneumonia in reference period |
| | Percent of children with diarrhea receiving ORT | Children under five with diarrhea in reference period |
| | Vitamin A coverage | Children 6-59 months |
| | Percent of infants exclusively breastfed through age 6 months | Children 0-5 months |
| | Percent of households treating water effectively | Household |
| FP | Modern contraceptive prevalence rate | Women in union 15-49 |
| | Percent of need satisfied by modern methods of family planning | Women in union 15-49, fecund, wanting to space or limit |
| | Percent of births spaced 3 or more years apart | Closed birth intervals during reference period |
| | Early childbearing (% women who had a first birth before age 18) | Women 18-24 |
| | Percent of births of order 5 or higher | Births during reference period |

All else being equal, the precision of measurement of a binomial population depends on sample size and the value of the estimate. The larger the number of samples, the more precise the measure, and the closer to 50% the value of the estimate, the less precise the measure. For example, a particular indicator is found to have a value of 50% based on a sample of 200 individuals: the precision of that estimate is ± 10 percentage points[3]. If the level of precision required is ± 5% and everything else remains the same, the sample size must approach 1,000. On the other hand, if the value of the indicator is 80% and all else remains the same, that sample of 200 individuals will give a precision of ± 7.8%.

<u>How many households are needed?</u> The likelihood of finding an appropriate case in the household depends on the particular indicator. Table 2 groups the indicators by their denominators and presents data from recent surveys in Bangladesh and Kenya. Assuming that 200 cases are needed for each indicator, the number of households needed in Bangladesh would range from 200 for indicators based on the household itself to more than 6,000 for indicators based on children with certain characteristics[4]. In Kenya, with higher fertility rates, the number of households would range from 200 to 5,000.

**Table 2a.** Denominator Units per Household: Bangladesh – 2005 NSDP Evaluation Survey

| **Denominator Unit** | **Units per household** | **Households needed for N=200** | **Indicator** |
|---|---|---|---|
| Household | 1 | 200 | Households with at least one ITN |
| | | | Percent of households targeted for indoor residual spraying that have been sprayed |
| | | | Percent of households treating water effectively |
| Live births during reference period (use last 3 years)[1] | Rural: 0.322<br>Urban: 0.281 | 621<br>712 | Percent of pregnant women who received two or more doses of SP for IPT |
| | | | Percent of women making greater than or equal to 4 antenatal care visits |
| | | | Percent of deliveries attended by skilled attendants |
| | | | Percent of births of order 5 or higher |
| Closed birth intervals during reference period (last 3 years) | [not collected] | | Percent of births spaced 3 or more years apart |
| Children under five | Rural: 0.601<br>Urban: 0.531 | 333<br>377 | Percent of children under five who slept under an ITN the previous night |
| | | | Percent of children underweight for age |
| Children 12-23 months | Rural: 0.111<br>Urban: 0.100 | 1,801<br>2,000 | Immunization coverage (DPT3) |
| Children 6-59 months[2] | Rural: 0.364<br>Urban: 0.355 | 549<br>563 | Vitamin A coverage |
| Children 0-5 months | Rural: 0.047<br>Urban: 0.037 | 4,255<br>5,405 | Percent of infants exclusively breastfed through age 6 months |
| Children under five with fever in reference period[4] | Rural: 0.166<br>Urban: 0.186 | 1,187<br>1,077 | Percent of children under five with suspected malaria receiving treatment within 24 hours |
| Children under five with symptoms of pneumonia | [not collected] | | Appropriate care seeking for pneumonia |

[3] 95% confidence interval and design effect of 2

[4] The number of households needed also depends on whether all index cases in the household are included in the survey or only one case. Fewer households are needed if all index cases are included.

**Table 2a.** Denominator Units per Household: Bangladesh – 2005 NSDP Evaluation Survey

| Denominator Unit | Units per household | Households needed for N=200 | Indicator |
|---|---|---|---|
| Children under five with diarrhea in reference period | Rural: 0.031<br>Urban: 0.036 | 6,430<br>5,459 | Percent of children with diarrhea receiving ORT |
| Pregnant women and children under five | Rural: 0.625<br>Urban: 0.552 | 320<br>362 | Percent of vulnerable populations protected by ITNs or indoor residual spraying |
| Women in union 15-49[5] | Rural: 0.8956<br>Urban: 0.9731 | 223<br>206 | Modern contraceptive prevalence rate |
| Women in union 15-49, who are fecund and want to space or limit | [not collected] | | Percent of need satisfied by modern methods of family planning |
| Women 18-24 | [not collected] | | Early childbearing |

1: Based on number of women with a birth in the last 3 years. We used only the last birth.
2: Children 9-59 months (most recent birth) for vitamin A coverage
4: Calculated for ARI, two younger children with fever in last two weeks.
5: Survey limited to ever-married women

**Table 2b.** Denominator Units per Household: Kenya – 2003 DHS

| Denominator Unit | Units per household | Households needed for N=200 | Indicator |
|---|---|---|---|
| Household | 1 | 200 | Households with at least one ITN |
| | | | Percent of households targeted for indoor residual spraying that have been sprayed |
| | | | Percent of households treating water effectively |
| Live births during reference period (last 5 years) | Rural: 0.87<br>Urban: 0.38 | 229<br>522 | Percent of pregnant women who received two or more doses of SP for IPT |
| | | | Percent of women making greater than or equal to 4 antenatal care visits |
| | | | Percent of deliveries attended by skilled attendants |
| | | | Percent of births of order 5 or higher |
| Closed birth intervals during reference period (last 5 years) | Rural: 0.68<br>Urban: 0.25 | 352<br>714 | Percent of births spaced 3 or more years apart |
| Children under five | Rural: 0.86<br>Urban: 0.35 | 232<br>574 | Percent of children under five who slept under an ITN the previous night |
| | | | Percent of children underweight for age |
| Children 12-23 months | Rural: 0.16<br>Urban: 0.07 | 1,216<br>2,998 | Immunization coverage (DPT3) |
| Children 6-59 months | Rural: 0.71<br>Urban: 0.32 | 284<br>631 | Vitamin A coverage |
| Children 0-5 months | Rural: 0.09<br>Urban: 0.04 | 2,304<br>5,056 | Percent of infants exclusively breastfed through age 6 months |

**Table 2b.** Denominator Units per Household: Kenya – 2003 DHS

| Denominator Unit | Units per household | Households needed for N=200 | Indicator |
|---|---|---|---|
| Children under five with fever in reference period | Rural: 0.33<br>Urban: 0.14 | 602<br>1,391 | Percent of children under five with suspected malaria receiving treatment within 24 hours |
| Children under five with symptoms of pneumonia | Rural: 0.36<br>Urban: 0.15 | 557<br>1,294 | Appropriate care seeking for pneumonia |
| Children under five with diarrhea in reference period | Rural: 0.12<br>Urban: 0.06 | 1,601<br>3,314 | Percent of children with diarrhea receiving ORT |
| Pregnant women and children under five | Rural: 0.95<br>Urban: 0.39 | 210<br>509 | Percent of vulnerable populations protected by ITNs or indoor residual spraying |
| Women in union 15-49 | Rural: 0.68<br>Urban: 0.37 | 296<br>547 | Modern contraceptive prevalence rate |
| Women in union 15-49, fecund, wanting to space or limit | Rural: 0.44<br>Urban: 0.24 | 450<br>827 | Percent of need satisfied by modern methods of family planning |
| Women 18-24 | [not in published report] | | Early childbearing |

Tables 2a and 2b demonstrate the minimum number of households needed to yield 200 index cases for the various health outcome indicators in two developing countries with different fertility levels. With a sample of 200 cases, the most conservative confidence interval around the estimates would be ± 10%, and many estimates will be more precise. However, it should be kept in mind that many outcome indicators change more slowly (for example, the upper limit for sustained growth in contraceptive prevalence appears to be 2-3% per year and many countries have shown lower growth). ***Country programs will have to trade off precision of the estimate (i.e. ability to detect small annual changes) with measurement costs***, as well as the relevance and importance to local program capacity-building.

## Background and experience to date with alternative data collection methodologies

Multi-stage sample surveys such as the Demographic and Health Survey (DHS) are a ‘gold standard’ for tracking health impact indicators such as birth and mortality rates at the national and subnational levels. However, the varying complexity and associated costs and time needed for execution and analysis might make these surveys less appropriate for more frequent outcome monitoring, and unless sampling frames are adjusted (such as oversampling target areas), national surveys may not be able to disaggregate findings down to the program level.

Over the last few decades, a number of smaller and more focused data collection approaches have been designed for routine health outcome monitoring of specified program interventions. In this section we describe three approaches to focused data collection: cluster/stratification surveys, rider/omnibus surveys, and longitudinal community surveillance.

1. **Cluster/stratification surveys**

- 30 by [n] survey (30 by 7, 30 by 10)

This two-stage cluster design was first developed by the WHO in 1978 and designed to estimate immunization coverage within ± 10 percentage points with 95% confidence (see earlier section on indicators and measurement). Sampling of clusters at the first stage is based on probability proportional to size (PPS). Therefore, if 30 clusters and then 7 subjects within each cluster are selected the sample is self-weighting. If a greater, but equal, number of subjects are sampled in each cluster, the sample is still self-weighting but precision of estimate is increased. Example: EPI survey (numerous implementations can be found at the WHO website (www.who.int).

This methodology has been adapted for USAID’s Child Survival and Health Grants Program. (CSHGP); the Child Survival Technical Support (CSTS) project provides technical assistance to PVOs who implement Knowledge, Practice and Coverage (KPC) surveys as part of their grants. In the 30 x10 surveys, 10 households are interviewed in each of the 30 clusters for a total of sample size of 300 and information on only one child under 2 years old is taken for each household.

- Lot Quality Assurance Survey (LQAS)

LQAS originated in the manufacturing industry for Quality Control purposes to assess quality of product meeting industry standards. Products were divided into lots and a statistically determined sample size was taken from the lot. If a certain number (based on production standard) of sampled products did not meet the quality standards, the whole lot was rejected. Thus the outcome is binary, indicating an acceptable or unacceptable lot. The original methodology has been modified over time to incorporate random sampling and double sampling schemes, which permits calculation of proportions of an attribute in the population (treated as a stratified sample), rather than just proportions of acceptability in lots.

LQAS application in public health started in the early 1980’s. LQAS has been used throughout the world for HIV/AIDS and STIs, immunization coverage, women’s and children’s health, growth and nutrition and diarrheal disease control[5]. Many developing country NGOs have applied LQAS for

[5] See Robertson, S and Valedez, J. (2006) Global Review of Health care surveys using Lot Quality Sampling 1984-2004, *Social Science and Medicine*, **63**: 1648-1660.

their projects, and international organizations, such as WHO, UNICEF, World Bank and USAID, have also applied this methodology for the monitoring of program outputs at local level.

- Cluster Sampling with Stratification

These surveys can cover a wide range of complexity. Since many decisions are made in the sample design, they are difficult to adequately describe briefly, and beyond the scope of this document. These surveys can include stratification, weights that may or may not differ from subject to subject sampled. Sample sizes can range from a few hundred to several thousand, depending on whether they were designed to measure few indicators or multiple indicators from several subpopulations (see earlier section on sample size and indicators). Often these are nationally representative stratified two-stage cluster designs. Cluster sampling can be adapted to a smaller size samples or a sub-national survey with considerably reduced implementing time and cost; however, it is more commonly known for larger national surveys. Examples: Demographic Health Surveys (DHS), Reproductive Health Surveys (RHS), Living Standard Measurement Survey (LSMS), Key Indicators Survey (KIS).

### 2. Rider survey/omnibus survey

The term “rider survey” refers to an additional data collection process “piggy-backed” onto a larger household survey effort. Rider surveys rely on, and take advantage of cost savings associated with, shared sampling, fieldwork, and processing of another, typically nationally representative survey, such as a quarterly employment or income and expenditure survey. Rider survey interviews are conducted either immediately following or within a matter of days after the host survey interview, and data processing is typically undertaken in conjunction with that of the host survey.

Rider surveys have been implemented in Ukraine, Jordan and the Philippines, to collect information for family planning and elements of MCH. Estimation of indicators at national, regional, and local levels, as well as for specific population strata, depends on sample size as well as on the representativeness of the sample. One of the advantages of the rider survey solution over the more comprehensive survey is the fact that cost savings associated with asking a smaller number of questions can be used to expand sample size to provide estimates for lower administrative levels. In the Philippines, sample sizes of the FP rider surveys have generally been larger than those of the DHS; those of the MCH rider surveys, smaller.

Omnibus surveys are typically conducted by commercial marketing research companies, usually in urban and peri-urban markets and stratified by socio-economic status; most companies conduct their surveys several times a year. A core set of indicators (e.g. age, sex, SES) is provided and the client is charged by the question for additional items. Omnibus surveys have been used by social marketing projects to collect family planning indicators; they are less likely to be found in rural areas.

### 3. Longitudinal Demographic and Community Surveillance

Community surveillance can serve as an alternative source for the longitudinal generation of specific outcome indicator packages in its own right and simultaneously as a platform for fielding periodic surveys or survey modules. Longitudinal community surveillance (LCS) carried out in demographic surveillance sites (DSS) has generally been conducted in locations or sentinel districts, often selected because of high prevalence of a disease or condition for which community interventions are being trialed and evaluated.

LCS/DSS has produced a full range of health indicators appropriate for malaria, MCH and family planning. There is a range of documented experience from LCS, including Tanzania, Vietnam, Indonesia, China and India. With multiple LCS sites in a single country, a broader longitudinal picture of conditions can be developed; a nationally representative set of surveillance points can provide data for measuring key outcomes.

### Cross-methodology comparison

Table 3 summarizes technical aspects associated with the three categories of data collection. Once the topics are fixed and the questionnaire designed, there are no major differences in ease of implementation among the data collection methodologies. As the different applications have evolved, especially within the variants of cluster sampling, the distinctions between cluster sampling and LQAS have become blurred. For example, more clusters and population units can be added to 30 by *n* cluster designs for more precision and they can use non-PPS design and integrated with stratification. LQAS can be adapted to include stratification with random sampling within strata (lots=strata in the case of LQAS) and complexity may be increased. One clear difference between rider surveys and the other data collection approaches is that rider surveys are not stand-alone but require a host survey to piggyback upon.

Costs of the each approach vary depending on size and complexity of the sample. Overall, rider surveys tend to be less expensive, but their timing depends on that of the host survey. Longitudinal community surveillance when designed as part an intervention program can be cost-shared with that program.

Table 3. Side-by-side comparison of selected methodologies

| **Method** | **Data Collection Method** | | | | |
|---|---|---|---|---|---|
| | **Cluster/Stratification Surveys** | | | **Rider Surveys** | **Longitudinal Community Surveillance** |
| | **"30 by *n*" Cluster Sampling** | **Lot Quality Assurance Sampling (LQAS)** | **Other Cluster Sampling Methods** | | |
| **Sampling Design** | Nationally representative two-stage cluster design with probability [of cluster selection] proportional to size (PPS) | Nationally representative stratified random sample design; knowledge needed of geographic boundaries of program areas | Nationally representative stratified two-stage (or more) cluster design | Usually attached to nationally representative survey of complex multi-stage design | Targeted sampling (e.g. to program implementation areas) or representative clusters |
| **Sub-populations** | Clusters based upon non-overlapping geographic boundaries | Lots based upon non-overlapping geographic boundaries | May incorporate strata containing clusters, clusters randomly selected, population elements randomly selected from clusters | Related to survey "piggybacked" upon; analysis population may be subset of main survey population | Usually defined by districts or smaller administrative units within targeted or sampled areas. |
| **Sample Size** | Multiples of 30 (e.g. 210, 300). Sample size must be adjusted based upon outcome measured, desired level of precision and sample design. | Dependent on the lot size, and thresholds. Sample size must be adjusted based upon outcome measured, desired level of precision and sample design. | Sample size must be adjusted based upon outcome measured, desired level of precision and sample design. | Variable; depends upon routine host survey | Typically >20,000 population, but variable, depending on indicators required. |

| **Method** | **Data Collection Method** | | | | |
|---|---|---|---|---|---|
| | **Cluster/Stratification Surveys** | | | **Rider Surveys** | **Longitudinal Community Surveillance** |
| | **“30 by *n*” Cluster Sampling** | **Lot Quality Assurance Sampling (LQAS)** | **Other Cluster Sampling Methods** | | |
| **Sampling Frame** | Sampling frame required identifying all clusters, and further detail only within each cluster selected at the first stage. | Sampling frame required to identify each lot, and detail within each lot | Sampling frame required identifying all clusters, and further detail within each cluster. | Likely to be derived from frame of routine host survey | Usually required, can rely upon census data (if available), maps. Relies on local information resources at smaller administrative level. |
| **Basis for Inference** | Confidence interval on proportion | Null hypothesis is that the lot is unacceptable. Proportions and CI are calculable | CI of survey estimates | CI of survey estimates | CI of surveillance/ rider survey estimates |
| **Outcome** | Overall estimate of proportion for all 30 clusters, but not at the level of each cluster | Individual lots (or program areas) judged acceptable or not, overall estimate of proportion for all lots can be calculated | Will provide estimates at lower levels if incorporated in sample design | Will provide estimates at lower levels if incorporated in sample design | Depends on sampling; stratum-level estimates possible. |
| **Precision** | ± 10 percentage points with sample size of 210, but can be more precise by increasing sample | Depends on design parameters, can be comparable to “30 by” with similar sample size. | Typically very high if properly designed. Low sample size applications are comparable to “30 by” and LQAS | Typically very high if routine host survey properly designed | Typically very high if properly designed |
| **Weighting** | Most commonly self-weighting. May use weights. | Weights can vary between lots | Each population element can have different weight, depending on design. | Depends upon host survey | Weighting has been used several times to aggregate LCS-site level data |

| **Method** | **Data Collection Method** | | | | |
|---|---|---|---|---|---|
| | **Cluster/Stratification Surveys** | | | **Rider Surveys** | **Longitudinal Community Surveillance** |
| | **“30 by *n*” Cluster Sampling** | **Lot Quality Assurance Sampling (LQAS)** | **Other Cluster Sampling Methods** | | |
| **Main Reasons for Potential Biases** | Homogeneity within clusters, non-random selection of population elements, non-sampling errors including non-response bias. | Small lot sizes, non-sampling errors including non-response bias | Non-sampling errors | Non-sampling errors | Low sampling biases within-sites if large enough sample; strong selection bias if generalizations made from targeted LCS sites. |
| **Level of Disaggregation** | National, district, smaller area depending on design. Not intended for cluster-level calculations | National and lot. Intended for program-level calculations. Can be simplified if only national-level estimates required. | National and at stratum-level calculations. Can be simplified if only national-level estimates required. | Depends on sample size as well as on the representativeness of the sample. | Community level possible, or sub-districts |
| **Typical Use** | Immunization, MCH,FP and HIV/AIDS rates | Operations research, immunization rates, HIV risk factors, MCH, post-disasters, quality management | Widely used for multiple indicators | Family planning use, prevalence and contraceptive security. Willingness to pay for FP. | Demographic and health surveillance; evaluation studies; health worker training in M&E |
| **Sampling Units** | Clusters and households | Depends upon the information required: program areas or households, or randomly selected individuals | Depends upon the type of information required. Clusters, and population elements or households | Clusters and households | Census enumeration areas (EA) within target administrative units, households |

## Evaluation criteria

The preceding sections described technical aspects of alternative data collection methodologies. Selection of a method or methods for field testing should be guided by these and other considerations. For the purpose of discussion we have grouped these into three inter-related dimensions: *utility, time and resources* and *management and coordination considerations*. Each dimension includes several components, which are briefly described below.

### Utility and Relevance to Missions

- Standardized across countries or sub-national focal areas: This criterion refers to cross-country and/or cross-site comparability. It is absolutely essential. Not only must each indicator measure the same thing in every country and every local context, but either the same data collection method must be adaptable across countries, or if different data collection methods are used, they must yield comparable findings.

- Valid and reliable measures of program outcomes: Validity has more to do with indicator selection. Reliability of the estimate is a key criterion for selection of the data collection method.

- Sensitive to short-term changes/ Precision of the estimates: There is an emerging consensus that outcome indicators could be measured every two years rather than annually. Even so, two years is a short time for some outcomes to be achieved, especially if it is expected that moderate changes in indicator point estimates are to be measured with any degree of statistical significance. Greater sensitivity to short-term changes requires more precise estimates, which in turn has implications for both sample design and management and costs (the second dimension).

- Level of aggregation and disaggregation: In many countries USAID assistance is not spread nation-wide but rather concentrated in selected geographic areas. In some countries, target areas may be larger sub-national units such as regions or provinces; in other countries target areas may be districts or even communities. Ideally, the data collection method would be suited for application at various levels, and if smaller geographic areas are selected for sampling, it would be advantageous if the small area estimates could be aggregated to large areas.

- Relevant and useful for decision-making at all levels (e.g., community, district, sub-region, national, international) and for many partners: M&E should be more than a program report card. Country Missions will be asked to cover the costs of data collection and analysis out of their operating budgets, so it is reasonable for them to expect that relevant and timely information generated can help inform their program and policy decisions. Thus one of the primary uses for Missions will be to monitor their PMP outcomes and SO indicators in priority program areas. The information generated should also be useful for other international donors and USAID CAs. Additional assistance may be required to develop capacity for information utilization.

- Capacity-building for relevance and transfer to host country counterparts: Host-country institutions should be central to field-testing and implementation from the outset. In the long term, host country counterparts should be capable of taking over data collection and analysis. This development perspective may be less important than other evaluation criteria in the short term.

**Time and resource considerations**

- In-country institutional and human capacity, need for external TA: Any methods or systems proposed for wide-scale implementation should not exceed the capacity of host country institutions to implement, manage and sustain within a time horizon of more than a few years. All else being equal, preference would be given to a data collection method that could rely on existing organizational and human capacity in the country with less need for external technical assistance, and/or to a new data collection method that could be mastered easily with minimal external assistance.

- Time needed for data collection and analysis: From the PEPFAR experience, we can expect that reporting deadlines will be rigorously enforced and firmly adhered to. Preference would be given to a data collection method with shorter turn-around time (e.g., 1-2 months after first trial year) from design to implementation to report production.

- Local implementation costs: As described earlier, country Missions will be expected to cover at least local costs (there may be limited central funding available for pilot testing and/or external assistance at start-up). A "reasonable" range of $100,000-$200,000 has been suggested[6].

**Management and coordination considerations**

- Opportunity costs/interference with service provision: Who will be expected to conduct data collection? Using service providers to collect data may increase the likelihood that the information will be used in program decision-making, but at the same time take time away from needed services.

- Flexibility for timeliness; Not only would preference be given to a data collection method with shorter turn-around time, but data collection itself must be scheduled to coincide with the USAID reporting cycle.

Table 4. Scoring matrix for sampling methodologies.

| Component | Points |
|---|---|
| **Utility** | |
| Standardized across countries and rolled up centrally | |
| Valid and reliable measures of program outcomes | |
| Sensitive to short-term changes/precision of the estimates | |
| Level of aggregation and disaggregation | |
| Relevant and useful for decision-making at all levels and for many partners | |
| Participatory and transferable to host country partners | |
| **Time and resource considerations** | |
| In-country institutional and human capacity, need for external TA | |
| Time needed for data collection, analysis and information transfer | |
| Local implementation costs ($) | |
| **Management and coordination considerations** | |
| Opportunity costs/non-interference with service provision; support of local M&E systems | |
| Flexibility for timeliness | |
| **Total** | **100** |

[6] Al Bartlett, personal communication.

## Designing data collection to measure effectiveness and impact of USAID assistance

The ideal M&E system would not only to track changes in outputs and outcomes over time, but attribute those changes to USG investment. In practice, there are considerable challenges to both reliable measurement of change and valid attribution of change to foreign assistance.

### Measurement issues, with special attention to sub-national estimates

Reliably detecting change requires that the differences between measurements be larger than the within-measurement sampling error. The precision of the estimates can be improved by increasing sample size, but this implies increased costs and time. However, not all indicators are susceptible to rapid change; in general, indicators that depend more on program activities may change more rapidly than indicators that depend more on changing beneficiary behaviors. For example, a door-to-door vaccination campaign might quickly raise vaccination coverage because children can be vaccinated on the spot, while an oral rehydration initiative might take longer to reach comparable coverage levels because parents and not health workers administer the therapy and only if/when the child develops diarrhea. Furthermore, if the denominator for the indicator spans several years, there may be so much overlap in the year-to-year results that they become less useful for monitoring program progress. For example, percentage of all births that are high parity (fifth or higher): if the reference period is all births in the last three years, then two-thirds of the births in year x will be repeated in the following year and their contribution will outweigh the impact of the most recent program year.

Bringing measurement down to the subnational or program area presents its own challenges. On the one hand, it should be easier to detect program-level changes when the universe is restricted to the program area. On the other hand, small area analysis is often complicated by random year-to-year fluctuations unrelated to the intervention, making it difficult to detect consistent trends over the short term.

A final measurement consideration is that different indicators have different denominators. All else being equal, it will take fewer sampled households to yield an adequate denominator of children (e.g. percentage of children under age 5 who slept under an ITN the previous night) than to yield an adequate denominator of recent births (e.g. percentage of women making 4 or more antenatal visits). Thus the precision of different indicators collected with the same survey will vary.

### Attribution issues

Substantiating causality (i.e. attributing improved outcomes to USG investment) implies demonstrating or at least accepting the counterfactual argument that if USG had not made the investment, the outcome would have had a different level than that observed. Therefore the ‘gold standard’ in evaluation design has often been field experiments with control or comparison areas or phased implementation of program interventions where the second set of areas to receive the intervention serves as control for the first areas to receive assistance. While such experimental designs are required for clinical trials and pilot-testing of new interventions, they are less feasible when the intervention is a scale-up of a proven activity or best practice.

To have impact at scale, USG funding usually interacts with the funds of host governments and other donors to support national or sub-national programs that deliver key interventions. For example, in a country where another donor such as UNICEF provides antibiotics, USG funding might support the planning and logistics required to deliver those commodities to community-based treatment programs for child illness. In another country, USG funding might support the development and training of primary and community health workers who diagnose child illness and dispense appropriate treatment. In a third country, USG funding might support the entire treatment program in one district while another donor supports the same program in the adjoining district. Output indicators for these programs would differ, but the key outcome indicator – sick children receiving antibiotic treatment – would be the same.

Effective integration of external assistance minimizes duplication of effort and maximizes impact of limited resources. At the same time, it complicates the demonstration of causality. In the latter case above where different geographic areas are 'assigned' to different development partners, there may be no non-intervention areas to serve as a control against which to measure the impact of USG investment. In the former cases where USG investment complements that of other development partners, the unique contribution of each partner is necessary but by itself not sufficient to produce the desired outcome.

Planning for evaluation of program impacts goes beyond the design of data collection instruments and procedures to the implementation of program interventions in the first place. 'After the fact' assessments can be conducted, but they are usually less satisfactory than if the evaluation design had been built into the program design at the planning stage. The 'three ones' approach to HIV/AIDS (one national plan, one national coordinating mechanism, one national evaluation plan) endorsed by the international community may be instructive for other health interventions as well.

The reality of program scale-up is that due to financial and human resource constraints not all operating units (districts, communities, etc.) will receive the health intervention at the same time; even within designated priority regions, some areas will start before others. This aspect of phased implementation can be exploited to provide experimental evidence of program impact. For example, consider a target district with 50 health facilities which will need rehabilitation, logistics support and staff training. Roll-out of the integrated package is scheduled to take 14 months (one facility per week). If facilities are randomly scheduled for roll-out[7] and outcomes measured at regular intervals in all 50 communities, then the facilities implementing the program later in the year or in year 2 can serve as a control for the first enrollees. This approach has been successfully implemented by the Progresa program in Mexico[8] and could be replicated by integrated health development assistance efforts if all partners work together.

If experimental designs are not feasible, there are a number of approaches for estimating causal effects from observational data. They include a range of econometric-based methods, such as difference-in-difference and instrumental variables models; as well as matching methods, such as, propensity score matching[9]. These methods require particular types of information and accepting specific assumptions for producing valid estimates of program effects.

[7] A stronger evaluation design than simple random selection would be matched groups (along one or more criteria known or thought to be related to the outcome indicators) with random within-group assignment.

[8] See Attanasio et al. (2001), "Educational Choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa". http://www.ifs.org.uk/edepo/wps/ewp0404.pdf.

[9] See Rubin, D.B. (1997), "Estimating causal effects from large data sets using propensity scores". *Annals of Internal Medicine*, **127**, 757-763. JHU Center for Communication Programs has used propensity scoring to measure the impact of communication programs when baseline data are unavailable.