Evaluation FAQ: **What Sample Size Do I Need for an Impact Evaluation?**

An appropriate study sample size and method of selection is critical to the success of any evaluation. In an impact evaluation, with a comparison drawn between a treatment group and a comparison group or showing a change over time, the following factors should be considered during the design phase. This will ensure that an adequate sample is selected to answer your primary evaluation questions.

- **Statistical significance**: This is the probability of committing a Type I error. A Type I error is when a change in a study indicator is detected, when in reality there has been no change. You need a larger sample size in order to have a lower statistical significance (offering a higher confidence level).
- **Statistical power**: This is the probability of not committing a Type II error. A Type II error is when a change in a study indictor is not detected, when in reality there has been a change. You need a larger sample size to have a higher statistical power.
- **Baseline values of the study indicator for the treatment and comparison groups**: These values can be in the form of either a mean or proportion.
- **Minimum change to detect in the study indicator**: This is the minimum detectable change in the study indicator that you hope to find as a measurable result of the intervention. You cannot expect to detect program impact smaller than the minimum detectable change, given a fixed sample size and other fixed design parameters. With pre-set parameters of statistical significance and statistical power, the detection of a smaller change would require a larger sample size.
- **Variance (or standard deviation)**: When the program impact is measured through difference or change in the study indicator, the variance (i.e., squared standard deviation) of the study indicator is required to calculate sample size. An estimate of the variance can be obtained from reliable, local data sources, when available. With pre-set parameters of statistical significance and statistical power, a larger variance of the study indicator will result in a larger required sample size.
- **Design effect**: This is the effect of the sample design used on the variance of the study indicator, given a fixed sample size. Simple random sampling (SRS) is used for comparison with other sampling designs. SRS is a sampling design whereby units are selected with equal probability of selection and without any clustering or stratification. A design effect smaller than 1 suggests that the sampling design used is more efficient than SRS, whereas a value greater than 1 suggests a less efficient design. Given the same fixed parameters (statistical significance and power), a larger sample size is needed if the design effect is greater than 1, as is typically the case for cluster samples commonly used for population-based surveys.
- **Domains**: The study groups for analysis constitute domains. Typical domains in impact evaluation are treatment and comparison groups, geographic areas, and other subpopulations. The sample size calculation should be performed at the domain level if your study aims to estimate program impact on a specific area or subpopulation (domain). This process will result in a separate sample size estimate for each domain. As such, the overall sample size is typically larger when you want to estimate impacts on different domains.
- **Number of respondents/participants per sampling unit**: The sampling units for a study sometimes differ from the observation unit of interest. For example, the observation unit of interest may be a woman of reproductive age (WRA) but the sampling unit to identify such women is the household. The sample size for units of observation required (e.g., WRA) has to be translated into the number of sampling units (e.g., households) based on the expected number of observation units per sampling unit (e.g., WRA per household).

USAID FROM THE AMERICAN PEOPLE PEPFAR MEASURE Evaluation

- **Response/participation rate**: For studies that rely on voluntary participation, study subjects may refuse to participate altogether (unit nonresponse) or may not answer some of the questions (item nonresponses). Nonresponse reduces the number of observations available for analysis. Sample size estimates need to be increased to compensate for expected nonresponse. An estimate of the potential nonresponse rate can be based on the nonresponse rates experienced in other studies of similar design, topic, and setting.

Often, there are multiple indicators of interest that require different sample sizes in one evaluation study. In such cases, the sample size is typically estimated based on the indicator that will require the largest sample—that is, the one with the most constraints. In this way, sample size requirements of all the other indicators will automatically be satisfied. Sometimes the sample size calculated, taking all these factors into account, is too large to be supported by the budget available for the evaluation. An option is to fix the sample size at the maximum your budget will support and then estimate the minimum detectable change (MDC) in your outcome indicator that is possible to measure with the sample size your budget supports. Essentially, this involves working backward through the sampling calculation to predict how much change you might be able to detect, given the constraints and assumptions you have built into the design. You will then need to decide whether the MDC is sufficient for the study to produce meaningful and useful results.

Table 1 provides examples of sample sizes from recent evaluation studies conducted by MEASURE Evaluation. This table is meant to illustrate the range of likely sample sizes based on study designs and outcomes. However, each impact evaluation requires sampling estimations based on the specifics of that evaluation.

Table 1: Examples of Sample Sizes for Impact Evaluations

Study	Sample	Comments
Bangladesh Smiling Sun Franchise baseline survey	34,300 Total: 7,300 urban/project 5,800 urban/non-project 14,000 rural/project 7,200 rural/non-project Unit: Households	A multistage, cluster-based household survey to measure change in two key maternal health indicators among women 10–49 years of age who had a birth in the three years preceding the survey. The sampling was designed to measure indicators comparing project and non-project areas in urban and rural domains.
Guatemala Feed the Future (FTF) baseline survey	6,301 Total: 1,264 project group 1 1,746 project group 2 997 project group 3 1,438 comparison group 1 856 comparison group 2 Unit: Households	A stratified multistage household survey to measure change in two key indicators related to nutritional status among children and poverty level. The sampling was designed to measure the indicators comparing project and non-project groups with different potential exposure to the program of interest. There were three project domains and two comparison domains.
Nigeria COMPASS program endline survey	4,500 Total (only project areas, no comparison group) Unit: Households	Endline household survey in five program local government areas (LGAs). Used a multistage stratified sampling strategy to select households. Based sample on estimates for two contraceptive use indicators and three immunization indicators.
Ukraine tuberculosis impact evaluation baseline data collection	Q1: 1,800 Total: 445 high-risk/project 445 high-risk/non-project 445 low-risk/project 445 low-risk/non-project Q2: 2,500 Total: 1,250 project and non-project Unit: Patient records	Q1: Impact of social support program on treatment default rates. Sample of medical records from selected facilities. Sampling allows for comparisons between high-risk and low-risk patients in intervention and comparison facilities. Indicator for sample size calculation was expected probability of treatment default. Q2: Impact of integrated services on tuberculosis and HIV care. Sample of medical records from selected facilities. Sampling allows for baseline collection in intervention and comparison facilities powered by testing rates and anitretroviral therapy initiation.
Jamaica randomized controlled trial for HIV prevention interventions	2,948 Total: 711 men/project 845 women/project 654 men/non-project 738 women/non-project Unit: Site patrons	Two-arm randomized controlled trial with stratification by gender. Site-based sampling of patrons. Primary outcome of interest was proportion of patrons with new sexual partners who report condom use.

www.measureevaluation.org measure@unc.edu

MEASURE Evaluation is funded by the U.S. Agency for International Development (USAID) under terms of Cooperative Agreement AID-OAA-L-14-00004 and implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International, John Snow, Inc., Management Sciences for Health, Palladium, and Tulane University. The views expressed in this presentation do not necessarily reflect the views of USAID or the United States government. FS-15-157