IMPACT EVALUATION IN PRACTICE

GLOSSARY

Activity. Actions taken or work performed through which inputs, such as funds, technical assistance and other types of resources are mobilized to produce specific outputs

Alternative hypothesis. In impact evaluation, the alternative hypothesis is usually the hypothesis that the null hypothesis is false; in other words, that the intervention has an impact on outcomes.

Attrition. When some units drop from the sample between one data collection round and another, for example because migrants are not tracked. Attrition is a case of unit non-response. Attrition can create bias in impact evaluations if it is are correlated with treatment status.

Baseline. Pre-intervention, ex-ante. The situation prior to an intervention, against which progress can be assessed or comparisons made. Baseline data are collected before a program or policy is implemented to assess the 'before' state.

Before and after comparison. Also known as pre-post comparison or reflexive comparison: A before and after comparison attempts to establish the impact of a program by tracking changes in outcomes for program beneficiaries over time using measures both before and after the program or policy is implemented.

Bias. The bias of an estimator is the difference between an estimator's expectation and the true value of the parameter being estimated. In impact evaluation, this is the difference between the impact that you calculate and the true impact of the program.

Census data. Data that covers all units in the population of interest (universe). *Contrast with:* survey data.

Cluster. A cluster is a group of units that are similar in one way or another. For example, in a sampling of school children, children who attend the same school would belong to a cluster, because they share the same school facilities and teachers and live in the same neighborhood.

Cluster sample. Sample obtained by drawing a random sample of clusters, after which either all units in selected clusters constitute the sample or a number of units within each selected cluster is randomly drawn. Each cluster has a well-defined probability of being selected, and units within a selected clusters also have a well-defined probability of being drawn.

Comparison group. Also known as control group. A valid comparison group will have the same characteristics as the group of beneficiaries of the program ("treatment group"), except for the fact that the units in the comparison group do not benefit from the program. Comparison groups are used to estimate the counterfactual.

Counterfactual. The counterfactual is an estimate of what the outcome (Y) would have been for a program participant in the absence of the program (P). By definition, the counterfactual cannot be observed. Therefore it must be estimated using comparison groups.

Cost-benefit. Ex-ante calculations of total expected costs and benefits used to appraise or assess project proposals. Cost-benefit can be calculated ex-post in impact evaluations if the benefits can be quantified in monetary terms and the cost data is available.

Cost-effectiveness. Entails comparing similar interventions based on cost and effectiveness. For example, impact evaluations on various education programs allow policy makers to make more informed decisions about which intervention may achieve the desired objectives given their particular context and constraints.

Difference-in-differences. Also known as double difference or D-in-D. Difference-in-differences estimates the counterfactual for the *change* in outcome for the treatment group by taking the *change* in outcome for the comparison group. This method allows us to take into account any differences between the treatment and comparison groups that are constant over time. The two differences are thus before and after and between the treatment and comparison groups.

Effect. Intended or unintended change due directly or indirectly to an intervention

Estimator. In statistics, an estimator is a statistic (a function of the observable sample data) that is used to estimate an unknown population parameter; an **estimate** is the result from the actual application of the function to a particular sample of data.

Evaluation. Evaluations are periodic, objective assessments of a planned, ongoing or completed project, program, or policy. Evaluations are used to answer specific questions often related to design, implementation and/or results.

External validity. External validity means that the causal impact discovered in the impact evaluation can be generalized to the universe of all eligible units. For an evaluation to be externally valid, it is necessary that the evaluation sample be a representative sample of the universe of eligible units.

Follow-up survey. Also known as "post-intervention" or "ex-post" survey: a survey that is fielded after the program has started, once the beneficiaries have benefited from the program for some time. An impact evaluation can include several follow-up surveys.

Hawthorne effect. The "Hawthorne effect" occurs when the mere fact that you are observing units makes them behave differently.

Hypothesis: A **hypothesis** (from Greek *ὑπόθεσις*; plural hypotheses) is a proposed explanation for an observable phenomenon. See also: null hypothesis and alternative hypothesis.

Impact evaluation. An impact evaluation is an evaluation that tries to make a causal link between a program or intervention and a set of outcomes. An impact evaluation tries to answer the question of whether a program is responsible for changes in the outcomes of interest. *Contrast with*. Process evaluation.

Indicator. An indicator is a variable that measures a phenomenon of interest to the evaluator. The phenomenon can be an input, an output, an outcome, a characteristic or an attribute.

Inputs. The financial, human, and material resources used for the development intervention.

Instrumental variable. An instrumental variable is a variable that helps identify the causal impact of a program, when participation in the program is partly determined by the potential beneficiaries. A variable must have two characteristics to qualify as a good instrumental variable: (i) it must be correlated with program participation; and (ii) it may not be correlated with outcomes Y (apart from through program participation) or with unobserved variables.

Intra-cluster correlation. Intra-cluster correlation is correlation (or similarity) in outcomes or characteristics between units that belong to the same cluster. For example, children that attend the same school would typically be similar or correlated in terms of their area of residence or socio-economic background.

Internal validity. Internal validity means that the impact evaluation uses a valid comparison group, that is, a comparison group that is a valid estimate of the counterfactual. **Intention-to-treat or ITT estimator.** The ITT estimator is just the straight difference in the outcome indicator Y for the group to whom we offered treatment and the same indicator for the group to whom we did not offer treatment. *Contrast with:* Treatment-on-the-treated.

John Henry effect. The "John Henry effect" happens when comparison units work harder to compensate for not being offered a treatment. When one compares treated units to those "harder-working" comparison units, the estimate of the impact of the program will be biased: we will estimate a smaller impact of the program than the true impact we would find if the comparison units did not make the additional effort.

Matching. Matching is a non-experimental evaluation method that uses large datasets and heavy statistical techniques to construct the best possible comparison group for a given treatment group.

Minimum desired effect. Minimum change in outcomes that would justify the investment that has been made in an intervention, accounting not only on the cost of the program and the type of benefits that it provides, but also on the opportunity cost of not having investing funds in an alternative intervention. The minimum desired effect is an input for power calculations: evaluation samples need to be large enough to detect at least the minimum desired effects with sufficient power.

Monitoring. Monitoring is the continuous process of collecting and analyzing information in order to assess how well a project, program or policy, is performing. It relies primarily on administrative data to track performance against expected results, make comparisons across programs and analyze trends over time. Usually monitoring tracks inputs, activities, and outputs, though occasionally it includes outcomes as well. Monitoring is used to inform day-to-day management and decisions.

Non-response. Arise when data is missing or incomplete for some sampled units. Unit non-response arises when no information is available for some sample units, i.e. when the actual sample is different than the planned sample. Attrition constitutes one form of unit non-response. Item non-

response occurs when data is incomplete for some sampled units at a point in time. Non-response may cause bias in evaluation results if it is associated with treatment status.

Null hypothesis. A null hypothesis is a hypothesis that might be falsified on the basis of observed data. The null hypothesis typically proposes a general or default position. In impact evaluation, the default position is usually that there is no difference between the treatment and control group, or in other words, that the intervention has no impact on outcomes.

Outcome. Can be intermediate or final. An outcome is a result of interest that comes about through a combination of supply and demand factors. For example, if an intervention leads to more supply of vaccination services, then actual vaccination numbers would be an outcome, as they depend not only on the supply of vaccines but also on the behavior of the intended beneficiaries: do they show up at the service point to be vaccinated? Final/long-term outcomes are more distant outcomes; this distance can be interpreted in a time dimension (it takes a long time before one gets to the outcome), or a causal dimension (there are many causal links needed to reach the outcome).

Output. The products, capital goods and services which are produced (supplied) directly by an intervention. Outputs may also include changes that result from the intervention which are relevant to the achievement of outcomes.

Power. The power of a test is equal to one minus the probability of a type II error, ranging from 0 to 1. Popular levels of power are 0.8 and 0.9. High levels of power are more conservative and decrease the likelihood of a type II error. An impact evaluation has high power if there is a low risk of not detecting real program impacts, i.e. of committing a type II error.

Power Calculations. Power calculations indicate the sample size required for an evaluation to detect a given minimum desired effect. Power calculations depend on parameters such as power (or the likelihood of type II error), significance level, variance and intra-cluster correlation of the outcome of interest.

Process evaluation. A process evaluation is an evaluation that tries to establish the level of quality or success of the processes of a program: for example: adequacy of the administrative processes, acceptability of the program benefits, clarity of the information campaign, internal dynamics of implementing organizations, their policy instruments, their service delivery mechanisms, their management practices, and the linkages among these. *Contrast with*: Impact evaluation.

Random sample. The best way to avoid a biased or unrepresentative sample is to select a random sample. A random sample is a probability sample where each individual in the population being sampled has an equal chance (probability) of being selected.

Randomized assignment or randomized control designs. Randomized assignment is considered to be the most robust method for estimating counterfactuals and is often referred to as the "gold standard" of impact evaluation. With this method, beneficiaries are randomly selected to receive an intervention and each has an equal chance of receiving the program. With large enough sample sizes, the process of random assignment ensures equivalence, in both observed and unobserved characteristics, between the treatment and control groups, thereby addressing any selection bias.

Randomized offering. Randomized offering is a method for identifying the impact of an intervention. When the program administrator can randomly select the units to whom s/he can offer the treatment from the universe of eligible units, but s/he cannot obtain perfect compliance: s/he cannot force any unit to participate/accept the treatment, and s/he cannot refuse to let a unit participate if the unit insists on participating/being treated. In the randomized offering method, the randomized offering of the program is used as an instrumental variable for actual program participation.

Randomized promotion. Randomized promotion is a similar method to randomized offering. Instead of randomly selecting units to whom the treatment is offered, units are randomly selected for promotion of the treatment. In this way, the program is left open to every unit.

Randomized selection methods. Randomized selection method is a group name for several methods that use a random assignment to identify the counterfactual: among them are: randomized assignment of the treatment, randomized offering of the treatment, and randomized promotion.

Regression. In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. In impact evaluation, regression analysis helps us understand how the typical value of the outcome indicator Y (dependent variable) changes when the assignment to treatment or comparison group P (independent variable) is varied, while the characteristics of the beneficiaries (other independent variables) are held fixed.

Regression discontinuity design (RDD). Regression Discontinuity Design (RDD) is a non-experimental evaluation method. It is adequate for programs that use a continuous index to rank potential beneficiaries, and that have a threshold along the index that determines whether potential beneficiaries receive the program or not. The cutoff threshold for program eligibility provides a dividing point between the treatment and comparison groups.

Results Chain. The program logic that explains how the development objective is to be achieved. It shows the link from inputs to activities to outputs to results.

Sample. In statistics, a sample is a subset of a population. Typically, the population is very large, making a census or a complete enumeration of all the values in the population impractical or impossible. Instead, researchers can select a representative subset of the population (using a sampling frame) and collect statistics on the sample that may be used to make inferences or extrapolate to the population. This process is referred to as **sampling.**

Sampling. Process by which units are drawn from the sampling frame built from the population of interest (universe). Various alternative sampling procedures can be used. *Probability sampling* methods are the most rigorous since they assign a well-defined probability for each unit to be drawn. Random sampling, stratified random sampling and cluster sampling constitute probability sampling methods. Non-probabilistic sampling (such as purposive or convenience sampling) can create sampling errors.

Sampling frame: Most comprehensive list of units from the population of interest (universe) that can be obtained. Differences between the sampling frame and the population of interest create a

coverage (sampling) bias. In the presence of coverage bias, results from the sample do not have external validity for the entire population of interest.

Selection bias. Selection bias occurs when the reasons for which an individual participates in a program are correlated with outcomes. This bias commonly occurs when the comparison group is ineligible or self-selects out of treatment.

Significance level. The significance level is usually denoted by the Greek symbol, α (alpha). Popular levels of significance are 5% (0.05), 1% (0.01) and 0.1% (0.001). If a test of significance gives a p-value lower than the α-level, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. The lower the significance level, the stronger the evidence required. Choosing level of significance is an arbitrary task, but for many applications, a level of 5% is chosen, for no better reason than that it is conventional.

Spillover effect. Also known as contamination of the comparison group. A spillover effect happens when the comparison group is affected by the treatment administered to the treatment group, even though the treatment is not administered directly to the comparison group. If the spillover effect on the comparison group is negative (i.e. they suffer because of the program), then the straight difference between outcomes in the treatment and comparison groups will yield an over-estimation of the program impact. By contrast, if the spillover effect on the comparison group is positive (i.e. they benefit), then it will yield an under-estimation of the program impact.

Statistical power. The **power** of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true (i.e. that it will not make a Type II error). As power increases, the chances of a Type II error decrease. The probability of a Type II error is referred to as the false negative rate (β). Therefore power is equal to $1 - \beta$.

Stratified sample. Obtained by dividing the population of interest (sampling frame) into groups (for example, male and female), then by drawing a random sample within each group. A stratified sample is a probabilistic sample: every unit in each group (or strata) has the same probability of being drawn.

Survey data. Data that covers a sample of the population of interest. *Contrast with:* census data.

Treatment-on-the-Treated (Effect of). Also known as TOT estimator. The effect of treatment-on-the-treated is the impact of the treatment on those units that have actually benefitted from the treatment. *Contrast with*: Intention-to-Treat.

Treatment group. Also known as the treated group or the intervention group. The treatment group is the group of units that benefits from an intervention versus the comparison group that does not.

Type I error. Error committed when rejecting a null hypothesis even though the null hypothesis actually holds. In the context of an impact evaluation, a type I error is made when an evaluation concludes that a program has had an impact (i.e. the null hypothesis of 'no impact' is rejected) even though in reality the program had no impact (i.e. the null hypothesis holds). The significance level determines the probability of committing a type I error.

Type II error. Error committed when accepting (not rejecting) the null hypohesis even though the null hypothesis does not hold. In the context of an impact evaluation, a type II error is made when concluding that a program has no impact (i.e. the null hypothesis of 'no impact' is not rejected) even though the program did have an impact (i.e. the null hypothesis does not hold). The probability of committing a type II error is 1 minus the power level.

Universe. The group of units that are eligible for receiving an intervention or treatment. The universe is sometimes called the population of interest.

Variable. In statistical terminology, a variable is a symbol that stands for a value that may vary.