



Impact Evaluation Glossary

Version No. 7

Date: July 2012

Please send comments or suggestions on this glossary to 3ie@3ieimpact.org.

Recommended citation: 3ie (2012) 3ie impact evaluation glossary. International Initiative for Impact Evaluation: New Delhi, India.

Attribution

The extent to which the observed change in *outcome* is the result of the *intervention*, having allowed for all other factors which may also affect the *outcome(s)* of interest.

Attrition

Either the drop out of *participants* from the *treatment group* during the *intervention*, or failure to collect data from a unit in subsequent rounds of a *panel data survey*. Either form of attrition can result in *biased* impact estimates.

Average treatment effect

The average value of the *impact* on the *beneficiary group* (or *treatment group*). See also *intention to treat* and *treatment of the treated*.

Baseline survey and baseline data

A survey to collect data prior to the start of the *intervention*. Baseline data are necessary to conduct *double difference* analysis, and should be collected from both *treatment* and *comparison* groups.

Before versus after

See *single difference*.

Beneficiary or beneficiaries

Beneficiaries are the individuals, firms, facilities, villages or similar that are exposed to an intervention with beneficial intentions.

Bias

The extent to which the estimate of *impact* differs from the true value as result of problems in the evaluation or sample design (i.e. not due to *sampling error*).

Blinding

A process of concealing which subjects are in the *treatment group* and which are in the *comparison group*, which is single-blinding. In a double blinded approach neither the subjects nor those conducting the trial know who is in which group, and in a triple blinded trial, those analyzing the data do not know which group is which. Blinding is generally not practical for socio-economic development interventions, thus introducing possible *bias*.

Cluster sample

A multi-stage *sample design*, in which a sample is first drawn of geographical areas (e.g. sub-districts or villages), and then a sample of households, firms, facilities or whatever, drawn from within the selected districts. The design results in larger standard errors than would occur in simple random sample, but is often used for reasons of cost.

Comparison Group

A group of individuals whose characteristics are similar to those of the treatment groups (or participants) but who do not receive the intervention. Under trial conditions in which the evaluator can ensure that no *confounding factors* affect the comparison group it is called a *control group*.

Confidence level

The level of certainty that the true value of *impact* (or any other statistical estimate) will be included within a specified range.

Confounding factors

Factors (variables) other than the programme which affect the *outcome* of interest.

Contamination

When members of the comparison group are affected by either the *intervention* (see *spillover effects*) or another intervention which also affects the *outcome* of interest. Contamination is a common problem as there are multiple development interventions in most communities.

Control Group

A special case of the *comparison group*, in which the evaluator can control the environment and so limit *confounding factors*.

Cost-benefit analysis (CBA)

A comparison of all the costs and benefits of the *intervention*, in which these costs and benefits are all assigned a monetary value. The advantage of CBA over analysis of *cost effectiveness*, is that it can cope with multiple outcomes, and allow comparison in the return to spending in different sectors (and so aid the efficient allocation of development resources).

Cost-effectiveness

An analysis of the cost of achieving a one unit change in the *outcome*. The advantage compared to *cost-benefit analysis*, is that the, often controversial, valuation of the outcome is avoided. Can be used to compare the relative efficiency of programs to achieve the outcome of interest.

Counterfactual

The state of the world in the absence of the *intervention*. For most impact evaluations the counterfactual is the value of the *outcome* for the *treatment group* in the absence of the *intervention*. However, studies should also pay attention to unintended outcomes, including effects on non-beneficiaries.

Dependent variable

A variable believed to be predicted by or caused by one or more other variables (*independent variables*). The term is commonly used in *regression analysis*.

Dichotomous variable

A variable with only two possible values, for example, "sex" (male=0, female = 1). The *dependent variable* in the *probit* participation equation estimated for *propensity score matching* is a dichotomous variable for which participate=1, didn't participate=0.

Difference-in-difference

See *double difference*.

Double difference

The difference in the change in the outcome observed in the treatment group compared to the change observed in the comparison group; or, equivalently, the change in the difference in the outcome between treatment and comparison. Double differencing removes selection bias resulting from time-invariant unobservables. Also called Difference-in-difference. Compare to single difference and triple difference.

Dummy Variables

A dichotomous variable commonly used in *regression analysis*. *Impact evaluation* often uses a dummy variable for program participation (participate=1, didn't participate=0) as an *independent variable* in a regression in which the *dependent variable* is the *outcome* variable.

Effect Size

The size of the relationship between two variables (particularly between program variables and outcomes). See also *minimum effect size*.

Eligible population

Those who meet the criteria to be *beneficiaries* of the *intervention*. The population may be individuals, facilities (e.g. schools or clinics), firms or whatever.

Encouragement design

A form of *randomized control trial* in which the treatment group is given an intervention (e.g. a financial incentive or information) to encourage them to participate in the intervention being evaluated. The population in both treatment and control have access to the intervention being evaluated, so the design is suitable for national-level policies and programmes.

Ex ante evaluation design

An impact evaluation design prepared before the intervention takes place. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment, and the collection of baseline data from both treatment and comparison groups. Also called prospective evaluation.

Ex post evaluation design

An *impact evaluation* design prepared once the *intervention* has started, and possibly been completed. Unless there was *random assignment* then a *quasi-experimental design* has to be used.

Experimental Design

See *Randomized Control Trial*.

External Validity

The extent to which the results of the *impact evaluation* apply to another time or place.

Facility survey

A *survey* of a *sample* of facilities (usually for health or education, but could apply to police stations, training facilities and so on) that aims to assess the level and quality of all elements required to provide services. The *unit of observation* is the facility, though data may also be collected on staff in a separate facility staff survey (e.g. a teacher survey). If a facility survey is conducted alongside a household survey it is important that the *survey instruments* include information so as households can be linked to the facilities they use for the purposes of data analysis.

Factorial design

A *randomized control trial* with multiple treatment arms, in which one arm receives treatment A, a second arm treatment B, and a third both treatments (A+B). There may also be a fourth no treatment *control group*.

Hypothesis

A specific statement regarding the relationship between two variables. In

an *impact evaluation* the hypothesis typically relates to the expected *impact* of the *intervention* on the *outcome*.

Impact

How an intervention alters the state of the world. *Impact evaluations* typically focus on the effect of the *intervention* on the *outcome* for the *beneficiary population*.

Impact evaluation

A study of the *attribution* of changes in the *outcome* to the *intervention*. Impact evaluations have either an *experimental* or *quasi-experimental* design.

Impact heterogeneity

The variation in *impact* as a result of differences in context, beneficiary characteristic or implementation of the *intervention*.

Independent Variable

A variable believed to cause changes in the dependent variable, usually applied in *regression* analysis.

Intention to treat estimate

The average treatment effect calculated across the whole *treatment group*, regardless of whether they actually participated in the intervention or not. Compare to *treatment of the treated*.

Internal Validity

The validity of the evaluation design, i.e. whether it adequately handles issues such as *sample selection* (to minimize selection bias), *spillovers*, *contagion*, and *impact heterogeneity*.

Intervention

The project, program or policy which is the subject of the *impact evaluation*.

Large n impact evaluation

Studies applying statistical means to construct a *counterfactual*, which requires a sufficiently large sample size (n) to ensure statistical *power*.

Logic model

Describes how a program should work, presenting the causal chain from inputs, through activities and outputs, to outcomes. While logic models present a theory about the expected program outcome, they do not demonstrate whether the program caused the observed outcome. A theory-based approach examines the assumptions underlying the links in the logic model.

Matching

A method utilized to create *comparison groups*, in which groups or

individuals are matched to those in the *treatment group* based on characteristics felt to be relevant to the *outcome(s)* of the *intervention*.

Meta-analysis

The systematic analysis of a set of existing evaluations of similar programs in order to draw general conclusions, develop support for hypotheses, and/or produce an estimate of overall program effects.

Minimum effect size

The smallest effect size the researcher deems necessary to detect in the *impact evaluation*. Used to perform the *power calculation* necessary to determine required *sample size*.

Mixed methods

The use of both quantitative and qualitative methods in an impact evaluation design. Sometimes called Q-squared or Q2.

N

Number of cases. Uppercase "N" refers to the number of cases in the population. Lower case "n" refers to the number of cases in the sample.

Outcome(s)

A variable, or variables, which measure the *impact* of the *intervention*.

Panel data and panel survey

Data collected through consecutive surveys in which observations are collected on the same sample of respondents in each round. Panel data may suffer from *attrition*, which can result in *bias*.

Participant

An individual, facility, firm, village or whatever receiving the *intervention*. Also known *treatment group*.

Pipeline approach

An *impact evaluation* design in which the *comparison group* are those who have not yet received the intervention, but who are scheduled to do so. The assumption is that there will be no *selection bias*, since both *treatment* and *comparison* groups are to receive the interventions. However, the quality of the *matching* should be checked, since later *participants* may differ from those treated earlier.

Power

The ability of a study to detect an *impact*. Conducting a *power calculation* is a crucial step in impact evaluation design,

Power calculation

A calculation of the sample required for the *impact evaluation*, which depends on the *minimum effect size* and required level of *confidence*.

Primary Data

Data collected by the researcher specifically for the research project.

Propensity Score Matching (PSM)

A *quasi-experimental design* for estimating the *impact* of an *intervention*. The outcomes for the *treatment group* are compared to those for a *comparison group*, where the latter is constructed through matching based on propensity scores. The propensity score is the probability of participating in the intervention, as given by a *probit regression* on observed characteristics. These characteristics must not be affected by the intervention. PSM hence allows matching on multiple characteristics, by summarizing these characteristics in a single figure (the propensity score).

Quasi-Experimental Design

Impact evaluation designs used to determine impact in the absence of a *control group* from an *experimental design*. Many quasi-experimental methods, e.g. *propensity score matching* and *regression discontinuity design*, create a *comparison group* using statistical procedures. The intention is to ensure that the characteristics of the *treatment* and *comparison groups* are identical in all respects, other than the intervention, as would be the case from an *experimental design*. Other, *regression-based* approaches, have an implicit *counterfactual*, controlling for *selection bias* and other *confounding factors* through statistical procedures.

Random assignment

An *intervention* design in which members of the *eligible population* are assigned at random to either the *treatment group* or the control group (i.e. *random assignment*). That is, whether someone is in the treatment or control group is solely a matter of chance, and not a function of any of their characteristics (either observed or unobserved).

Randomized Controlled Trial (RCT).

An *impact evaluation* design in which *random assignment* has been used to allocate the *intervention* amongst members of the *eligible population*. Since there should be no correlation between *participant* characteristics and the *outcome*, and differences in *outcome* between the treatment and control can be fully attributed to the intervention, i.e. there is no *selection bias*. However, RCTs may be subject to several types of *bias* and so need follow strict *protocols*. Also called *Experimental design*.

Regression Analysis

A statistical method which determines the association between the *dependent variable* and one or more *independent variables*.

Regression discontinuity design (RDD)

An *impact evaluation* design in which the *treatment* and *comparison* groups are identified as being those just either side of some threshold value of a variable. This variable may be a score or observed characteristic (e.g. age or land holding) used by program staff in determining the *eligible population*, or it may be a variable found to distinguish *participants* from non-participants through data analysis. RDD is an example of a *quasi-experimental design*.

Replication

Independent verification of study findings. Internal replication attempts to reproduce study findings using the same dataset, whilst external replication evaluates the same intervention in a different setting or at a different time. Internal replication may be pure replication, which uses the same data and model specification, or may test robustness to different model specifications, estimation methods and software.

Sample

A subset of the *population* being studied. The sample is drawn randomly from the *sampling frame*. In a simple random sample all elements in the frame have an equal probability of being selected, but usually more complex sampling designs are used, requiring the use of *sample weights* in analysis.

Sampling Frame

The complete list of the *population* of interest in the study. This is not necessarily the complete population of the country or area being studied, but is restricted to the eligible population, e.g. families with children under five, or female –headed households. For a *facility survey*, the sampling frame would be all facilities in the area of study. If a recent sampling frame is not available then one needs to be constructed through a field-based listing.

Secondary Data

Data that has been collected for another purpose, but may be reanalyzed in a subsequent study.

Selection Bias

Potential biases introduced into a study by the selection of different types of people into treatment and comparison groups. As a result, the outcome differences may potentially be explained as a result of pre-existing differences between the groups, rather than the treatment itself.

Sampling error

The error which occurs as estimates are used making data from a sample rather than the whole population.

Sample weights

A technique used to ensure that statistics generated from the *sample* are

representative of the underlying *population* from which the sample is drawn. Sample weights should normally be used, though there is debate as to what to do when using *propensity score matching*, this is an alternative weighting system.

Single difference

Either, the comparison in the outcome for the treatment group after the *intervention* to its *baseline* value (also called *before versus after*), or an *ex post* comparison in the outcome between the *treatment* and *control groups*. Compare to *double difference*.

Small n impact evaluation

The set of best available methods when n is too small to apply statistical approaches to constructing a *counterfactual*.

Spillover effects

When the *intervention* has an *impact* (either positive or negative) on units not in the treatment group. Ignoring spillover effects results in a *biased* impact estimate. If there are spillover effects then the group of *beneficiaries* is larger than the group of *participants*. When the spillover affects members of the *comparison group*, this is a special case of *contagion*.

Survey

The collection of information using (1) a pre-defined *sampling* strategy, and (2) a *survey instrument*. A survey may collect data from individuals, households or other units such as firms or schools (see *facility survey*).

Survey instrument

A pre-designed form (questionnaire) used to collect data during a *survey*. A survey will typically have more than one survey instrument, e.g. a household survey and a *facility survey*.

Systematic Review

A synthesis of the research evidence on a particular topic, such as the effectiveness of water supply and sanitation, obtained through an exhaustive literature search for all relevant studies using scientific strategies to minimize error associated with appraising the design and results of studies. A systematic review is more thorough than a literature review. It may use the statistical techniques of a *meta-analysis*, but need not necessarily do so.

Theory-based impact evaluation

A study design which combines a *counterfactual* analysis of impact with an analysis of the causal chain, which mostly draws on factual analysis.

Theory of change

Laying out the underlying causal chain linking inputs, activities, outputs and outcomes, and identifying the assumptions required to hold if the

intervention is to be successful. A theory of change is the starting point for *theory-based impact evaluation*.

Treatment group

The group of people, firms, facilities or whatever who receive the intervention. Also called *participants*.

Treatment of the treated

The treatment of the treated estimate is the *impact (average treatment effect)* only on those who actually received the *intervention*. Compare to *intention to treat*.

Triple difference

The comparative or differential impact on two groups, calculated as the difference in the double difference impact estimate for each group compared to a no treatment *comparison group*. A significant triple difference estimates demonstrates the presence of *impact heterogeneity*.

Unit of analysis

The class of elemental units that constitute the population and the units selected for measurement; also, the class of elemental units to which the measurements are generalized.

Unobservables

Characteristics which cannot be observed or measured. The presence of unobservables can cause selection *bias* in *quasi-experimental designs*, if these unobservables are correlated with both participation in the programme and the *outcome(s)* of interest.