Glossary
From *Running Randomized Evaluations: A Practical Guide*, by Rachel Glennerster and Kudzai Takavarasha

attrition: When data are missing because we are unable to measure the outcomes of some of the people in the sample.

allocation ratio or fraction: The proportion of people within a given sample randomly allocated to receive the program.

adverse selection: A situation in which those with a higher risk of a bad outcome are more likely to participate in a program than those with a low risk. Individuals know their risk but program implementers cannot tell the risk of individuals.

anticipation effect: When the comparison group changes behavior because they expect to receive the treatment later on (or in a rotation design where those in the current treatment group change their behavior because they know they are going to become the comparison group later).

before/after comparison: A comparison that measures how outcomes for program participants changed over time.

Bonferroni adjustment: Used to adjust the confidence intervals around our coefficients for the fact that we are testing several different hypotheses. One way to do this is to multiply the p-value by the number of tests.

baseline: Measurement before the start of a program that can serve as a comparison.

causal impact: Any changes in outcomes that are caused by a program: the difference between outcomes with the program and what those outcomes would have been in the absence of the program.

comparison group: A group that is used to identify what would have happened in the absence of the program. In a randomized evaluation it consists of those randomly chosen not to receive access to the program.

compliance: When participants adhere to their assigned treatment regime.

conditional cash transfer: A program that that offers cash, conditional on participants complying with specific socially desirable behaviors.

confidence interval: A range of values around the an estimated value (e.g. estimated effect size) within which the true value is likely to fall. The confidence interval depends on the significance level we choose. Thus for a significance level of 95 percent, we have a 95 percent confidence that the true effect size falls within the specified confidence interval.

cost-benefit analysis: An approach to comparing the costs and benefits of different programs in which all the different benefits of a program are translated onto one scale (usually a monetary scale), and then compared to the costs of the program.

cost-effectiveness analysis: An analysis that tells us the cost it takes to achieve a given impact on a particular indicator; used to compare different programs that have the same objective measured using the same indicator.

counterfactual: How the people in the program would have fared if our program had never been implemented, used to understand the causal impact of a program.

critical value: The value of the estimated value (e.g. treatment effect) which exactly corresponds to the significance level—i.e. anything above this is statistically significantly different from zero at the significance level (e.g. at the 95 percent level), anything below it is not.

cross section comparison (*see* **simple difference comparison**)

data mining: Looking for the result you want in the data until you find it.

defiers: Individuals or groups who don't take up because they are allocated to the treatment or do take up because they are allocated to the control group; i.e., those who defy the assignment made by the evaluators.

demand effects: (aka response bias) When participants change behavior in response to their perception of the evaluators' objective.

descriptive survey: A survey that describes the current situation but does not attempt to answer causal questions about why the situation is as we find it.

difference-in-difference: A research approach that combines a before/after comparison with a participant/nonparticipant comparison. It measures change in outcomes over time of the program participants relative to the change in outcomes of nonparticipants.

dummy variables: A variable that takes either the value zero or one; i.e. a special type of binary variable.

encouragement design: A research design in which both treatment and comparison groups have access to the program but some individuals or groups are randomly assigned to receive an encouragement to take up the program.

endline: Measurement at the end of the study.

exclusion restriction: Exclusion restriction is an assumption that must hold if any instrumental variable strategy including an encouragement randomization strategy is used. It states that the instrument can only affect the outcome though its effect on the instrument. In other words it can not have a direct effect on the outcome.

evaluation-driven effects: The evaluation itself, and the way in which it is administered or the way outcomes are measured changes the way people behave, for example Hawthorne and John Henry Effects.

exact matching: When each participant in an evaluation is matched with at least one nonparticipant who is identical on selected observable characteristics, i.e., characteristics for which we have data, like age and occupation. Participants who have no direct match are dropped from the analysis.

experimental protocol: An agreement that describes intentions for treatment allocation, program and evaluation implementation, and data collection logistics.

externalities: (*see* spillovers)

external validity: (aka generalizability) When the results of evaluation are valid in contexts other than those the experiment was conducted in.

exogenous shocks: Events that generate variation in conditions across a study area and are not correlated with any underlying condition or characteristic of the population they effect. They allow us to discern causal relationships.

F test: Statistical test of whether two numbers are significantly different from each other.

false positive (aka Type I error or alpha error): Occurs when we find a statistically significant effect even though the program did not actually have a treatment effect. We then wrongly infer that the program had an effect.

false zero (aka Type II error): Occurs when we fail to find a significant effect even though there truly is a treatment effect. We wrongly infer that the program does not work.

general equilibrium effects: Effects on outcomes (such as prices and wages) that are determined by the pressure to equalize forces (usually demand and supply) derived by interactions across many people.

generalizability: (*see* external validity)

Hawthorne effects: When the treatment group works harder than normal in response to being part of an evaluation. An example of an evaluation driven effect.

heterogeneous treatment effects: When the effect of the program is different for different subgroups in a population.

implicit association test: An experimental method that relies on the idea that respondents who more quickly pair two concepts in a rapid categorization task subconsciously associate those concepts more strongly. It is used to test prejudice.

intracluster correlation: A measure of how much more correlated those within a cluster are compared to those in different clusters. A key component of the design effect which is a measure of how much less precise our estimate is, for a given sample size, when we move from an individual level randomization to a group or cluster-level randomization.

intertemporal correlation: A measure of how correlated outcomes for individuals are over time.

imprecise zero: When we can neither rule out a large effect nor a zero effect because our confidence bands are very wide.

indicator: An observable metric used to measure outcomes.

internal validity: When an evaluation can measure the causal impact of the intervention.

institutional review boards (IRBs): Committees that review research proposals to ensure they comply with ethical guidelines and whose permission is required before research involving people (human subjects) can proceed.

instrument: The tool we use to measure indicators.

instrumental variable: A variable that does not suffer from selection bias and is correlated with the outcome variable that allows us to estimate the causal impact through one very specific channel.

John Henry effect: When the comparison changes behavior in response to being part of an evaluation.

matched random assignment: A randomization procedure in which two units are matched on a list of important characteristics and then one of them is randomly assigned to treatment and the other to comparison.

midline: Measure in the middle of a program.

minimum detectable effect (MDE): The smallest effect that a given evaluation design will be able to detect with a given probability. We choose our sample size to be able to achieve an MDE.

monotonicity assumption: An assumption, made when we use an encouragement design, that everyone must be affected by the encouragement in the same direction.

moral hazard: When being insured against a risk makes people more likely to undertake risky behavior.

multivariate regression: Individuals who received the program are compared with those who did not, and other factors that might explain differences in the outcomes are "controlled" for. In other words, we run a regression of our outcome indicator against treatment status and other indicators that might explain our outcome variable.

needs assessment: Research that carefully collects descriptive information, both qualitative and quantitative about problems that may exist and the needs of the population a program is designed to serve.

noncompliance: When the randomized treatment assignment is not followed, i.e. when those assigned to treatment end up not receiving the program or those assigned to the comparison group receive the program.

null hypothesis: Assumption that that the treatment effect is zero for all subjects. It is designated $H_{0.}$

one-sided test: A test in which one looks for either a positive impact or a negative impact but not both in the same test.

outcome: A change or impact caused by the program we are evaluating.

oversubscribed: When there are more people interested in participation than the program has resources to serve.

phase-in design: A research design in which some people are selected to enter into the program at different times.

power (*see* statistical power)

power function: An equation relating power to its determinants: (1) the level of statistical significance, (2) the minimum detectable effect size that practitioners and policy-makers care about, (3) the variance of the outcome of interest, (4) the proportion of units allocated to the treatment, and (5) the sample size and if doing a group level randomization (6) the size of the group and (7) the intracluster correlation.

pre-analysis plan: A plan which describes, ahead of time, how the data will be analyzed. Conducted to address concerns about data mining.

process evaluation: Research that uses data collection instruments to tell whether the program is being implemented as planned.

proof-of-concept evaluation: An evaluation which tests whether an approach *can* be effective in the best possible situation even if that is not the form in which the approach would be implemented when scaled up.

purchasing power party (PPP): The rate at which we could purchase a standard basket of goods in two countries.

randomized assignment: In which we take a pool of eligible units—persons, schools, villages, firms—and then assign those units to treatment and comparison groups by a random process such as a toss of a coin, a random number generator, or a lottery.

randomization cells: The divisions that we randomize our list of eligible units into. In a simple evaluation that measures the impact of a program, we randomize our list of eligible units into two cells: treatment and comparison. In more complicated evaluations, we may have to divide our units into more than two cells.

randomization device: The method we use to randomize our units. This can be mechanical (a coin, dice, or ball machine), a random number table, or a computer program with a random number generator.

random number table: A list of unique numbers that are randomly ordered.

random sampling: In random sampling, we take a population and randomly select units to create a group that is representative of the entire population. We can then measure characteristics in this group and infer what the characteristics of the entire population are.

regression discontinuity design: When a program has a strict eligibility cut off based on measurable criteria this evaluation design can be used to evaluate the program by comparing outcomes for those just above and below the eligibility cutoff.

research hypothesis: The research hypothesis says that the treatment effect is not zero; chance alone could not account for the observed difference.

residual variance: The variance in the outcome variables between people (or units) that cannot be explained by the program or any control variables (e.g., gender, age, et cetera.) we may use in our analysis.

respondent: The person, group of people, or administrative data set we interview, test, observe, or access to measure the indicators.

rotation design: A research design in which everyone needs to receive the program, but resources are too limited to treat everyone at once and thus groups take turns receiving the program.

sample size: The number of units on which data are collected in the evaluation.

selection: The process by which participants are determined.

selection bias: When our estimated effect is biased by our failure to take into account the tendency for those who get a program to be different from those who do not get a program. We attribute differences in outcomes to the program, when they are actually caused by nonprogram-related differences in those who self-selected or are selected for the program.

s**imple difference comparison** (aka cross-section comparison): Measures the difference between program participants and nonparticipants after the program is completed.

simple random assignment: When one takes the list of eligible units as one big pool, and randomly assigns them to different cells.

social desirability bias: the tendency of participants to give an answer to a question that is in line with social norms even if this does not accurately reflect their experience.

spillovers (aka externalities): A program's effects on those who are not in the program. Spillovers can take many forms and can be positive or negative.

standard deviation: a measure of dispersion from the mean in the underlying population from which we sample.

standard error: A measure of precision of our estimated effect size (the larger our standard error the less precise our estimate). There is a 95 percent chance that the true effect size is within ±2 standard errors of our estimated effect size. Formally, the standard error of a sample is equal to the standard deviation of the underlying population divided by the square root of the sample size.

statistical matching: Program participants are compared to a group of nonparticipants that is constructed by finding people whose observable characteristics (age, income, education, etc.) are similar to those in the treatment group.

statistical power: The likelihood that our experiment will be able to detect a treatment effect of a specific size. For a given minimum detectable effect size and level of significance, power is a measure of how precisely we will be able to measure the impact of the program.

stratification or stratified random sampling: An assignment method in which we first divide the pool of eligible units into strata or groups based on observable characteristics and then within each s tratum follow the procedure for random assignment.

subgroup: Any group of individuals in a sample with at least one common characteristic established prior to the start of the program and evaluation.

survey effects: In which being surveyed changes subsequent behavior of the treatment or comparison groups.

T-test: A statistical test which determines whether an estimated coefficient or effect is statistically different from a specified value (usually zero).

theory of change: A theory, set out at the beginning of a program, specifying the steps in the pathways through which the intervention(s) could lead to impact**.**

treatment density: Treatment density is the proportion of people within a unit (a school, a community, a marketplace) who receive access to the program (i.e. are included in the treatment).

treatment group: Those who were randomly picked to receive access to the program.

treatment lottery: A research design in which units (individuals, households, schools, etc.) are randomly assigned to the treatment and comparison groups. The treatment group is given access to the program, the comparison group is not.

true difference: The difference we would find (on average) if we ran the perfect experiment an infinite number of times.

true positive: We have a true positive when we find a statistically significant effect and there truly is a treatment effect. We correctly infer that the program works.

true zero: We have a true zero when we find no statistically significant effect and there truly is no treatment effect. We correctly infer that the program does not work.

undersubscribed: When program take-up is low and the program is serving fewer people than it has resources to cover.

variable: The numeric values of indicators.