EBITE RESOURCES

Glossary Of Research Terms

ANCOVA (Analysis of covariance). Like ANOVAs, ANCOVAs compare the outcome means of three or more groups. ANCOVAs, however, include additional predictor variables, not just an intervention variable. They may include, for example, gender, age, or race/ethnicity, allowing a look at intervention effects over and above the effects of those demographic factors. (See ANOVA, Mean, Regression, and *t*-test.)

ANOVA (Analysis of variance). ANOVA is a statistical procedure that compares three or more group means (average scores) to see if they are statistically significantly different. There are only two variables in an intervention study ANOVA: the variable with information on intervention group (with at least 3 categories) and the outcome. ANOVAs are useful for testing whether average scores on an outcome are better for a group who received an intervention compared to a control group and a treatment-as-usual group. Some EBIs you'll encounter in online repositories have been tested with ANOVAs. As with any means comparison, ANOVAs summarize results for the students in each condition. They do not give specific information on how many students benefited from the intervention or characteristics of those who benefited or did not benefit. (See ANCOVA, Mean, Regression, Treatment-as-usual, and *t*-test.)

Attrition. Attrition is the loss of sample members during an intervention study. Attrition means some intended recipients of an intervention did not receive it. Participants may drop out for many different reasons, some of which can affect conclusions made about the effectiveness of the intervention. For example, if teachers stopped implementing an intervention because they did not think it was helping their students, researchers would end up analyzing results only from students who were most responsive to the intervention—making it look more effective. Some online EBI repositories will report attrition rates in studies. (See Intent-to-treat.)

Baseline Equivalence. Baseline equivalence refers to how similar the groups assigned to conditions are before the intervention begins. We want students (or others) who receive the intervention to be as similar as possible to those who receive a different intervention (e.g., the usual services) or no intervention. The more similar the starting characteristics of study participants across conditions, the more confident we can be that differences in outcomes at the end of the study were due to the intervention. Random assignment to conditions is the best strategy to achieve baseline equivalence, but some other strategies can allow some degree of confidence in results. Typical characteristics that are examined at baseline are gender, race/ethnicity, pretest scores on the outcomes of interest, and any other characteristics that may be predictive of outcome scores.

Bias. Bias refers to inaccuracies of estimates obtained in analyses. Statistical analyses can only provide values that are *estimates,* based on data from a study sample, of *true* values (the actual real but unknowable effect) for the theoretical population of students like those in the study. If

an estimate of an intervention's effect on an outcome is biased, it may be higher or lower than the true value for the population represented by the sample. Often researchers do not know how biased their estimates are. (See Standard error.)

Causality/Causal Inference. Researchers want to claim that better outcomes in the intervention condition are *caused by* their intervention. If groups are different at the beginning of the study (e.g., in terms of demographics, scores on outcome variables, willingness to participate, etc.), researchers cannot claim their intervention caused differences in outcome scores. The differences could be at least partly due to initial group differences. Demonstrating causality, is a central issue in intervention research. Researchers' ability to infer causality is derived from study design. (See Design, Experimental design, Internal validity, and Random assignment.)

Comparison group. Technically, a comparison group in a study comprises participants who are *non-randomly* assigned to the condition that does not receive the intervention. Outcome scores from members of the comparison group (or comparison condition) will be compared at the end of the study to scores from participants in the experimental group or condition. Because members of a comparison group were not randomly assigned to that group, researchers generally cannot claim their intervention *caused* different outcome scores across groups. Participants in the comparison and experimental groups may have differed from the start. There are some statistical procedures that can increase confidence in causal inference in this situation, but we will not be studying them. Often, the terms control group and comparison group are used interchangeably. (See Causality, Condition, Control group, and Random assignment.)

Condition (in a study). Condition refers to the group participants are assigned to in an intervention study. Most intervention studies have an experimental condition—the one receiving the intervention—and a control or comparison condition that does not received the intervention. Other studies may have more groups, such as a treatment-as-usual group or a group getting another version of the intervention. (See Comparison group, Control group, Random assignment, and Treatment-as-usual.)

Control group. Technically, a control group is the *randomly* assigned group in an experimental study that does not receive the intervention. Outcome scores from members of the control group (or control condition) will be compared at the end of the study to scores from the experimental group or condition. Having a randomly assigned control group helps researchers claim causality—that their intervention *caused* better outcome scores. Often, the terms control group and comparison group are used interchangeably. (See Design, Experimental study, Quasi-experimental study, Random assignment, and Comparison group.)

Correlation. Researchers want to claim that better scores in the intervention condition are *caused by* their intervention. The design of many studies, however, makes it impossible to make that claim. In cross-sectional studies, for example, data are collected at one time point. It is not possible to infer causality when no time passed in the study. Therefore, researchers can

demonstrate only that two variables are *correlated* or *associated* with each other. Correlated means that when scores on one variable go up, scores on the other also tend to predictably go up (positive correlation) or down (negative correlation). Correlations can also be used in studies with multiple time points and conditions, however. In a study with internal validity, a correlation between the intervention and an outcome is evidence of effectiveness. (See Design and specific study designs.)

Cross-sectional design. Studies using a cross-sectional design collect data at only one time point. Scores for different groups can be compared, but no claims can be made that one variable (e.g., the intervention variable) caused differences in those scores across groups. (See Design and specific study designs.)

Design (of a study). The design of a study includes how participants are grouped and the timing and number of data collections. (See Cross-sectional, Experimental, Longitudinal, Pretest/Post-test, Quasi-experimental, Single-subject design.) For example, in an educational intervention study, students may be placed into two groups, one of which receives the intervention and one of which does not; or students may not be separated into groups at all. (See Condition, Comparison group, Control group, Treatment-as-usual, and Random assignment.) Data may be collected once; before and after the intervention; or multiple times during and after the intervention. (See Follow-up data collection, Pretest, Post-test, Longitudinal.)

Disaggregating data. Disaggregating data means breaking it down by groups of interest (e.g., by race/ethnicity, gender, grade level, school) to see separate scores for each of those groups. Disaggregating data allows school staff and researchers to identify group-specific needs and strengths, and potential targets of tier 1 and 2 interventions. Disaggregating data is critical for evaluating interventions. Some interventions may help some students but do nothing for others. When choosing EBIs, you'll want to know if students like yours benefited.

Effect Size (ES). Effect size refers to the magnitude of change in outcomes that can be attributed to an intervention. The most commonly used ES is a simple formula based on group means. The difference in means between two experimental conditions or between the pre- and post-test scores for one group, is divided by a standard deviation (one of three possible standard deviation formulas). The result is an ES that can be compared across interventions. Some researchers say interventions worth using should have an ES of at least .20; other say .40. Larger effect sizes are better. Analyses can also show if the ES of an intervention is higher for one group of students than others. (See Disaggregating data.) ESs can be calculated for correlations, regression values, means comparisons and other statistics. (See ANOVA, ANCOVA, Correlation, and Regression.) The What Works Clearinghouse uses ESs based on an "improvement index," which is based on the average change in percentile ranking of students in an intervention group versus students in the control or comparison group.

Equivalent groups. The best way to isolate the effects of an intervention on recipients is to start a study with conditions comprising participants who have the same characteristics,

experiences, scores on targeted concerns (academic performance, mental health) and other variables. Many characteristics of individuals can be measured and compared across groups, meaning researchers can test equivalence on some characteristics across non-randomly assigned groups. However, even if conditions are equivalent in terms of observable characteristics such as gender, age, reading scores, etc., they may be different in terms of hidden or unmeasured characteristics. Random assignment is the only way to be confident that groups are equivalent. (See Causality, Conditions, Experimental design, and Internal validity).

Every Student Succeeds Act (ESSA). ESSA is federal legislation that has been in effect since 2015 requiring that schools achieve equitable educational outcomes for populations of students that have historically had lower performance. It also mandates the use of evidence-based interventions and annual assessments that provide school-, district-, and state-level performance scores. ESSA lists four levels of evidence that EBIs must demonstrate--—strong, moderate, promising, and demonstrates a rationale. (See EBITE ESSA lessons and resources.)

Evidence. Evidence of intervention effectiveness comes from intervention studies. Study design and magnitude of effects are essential to claims that there is evidence of an intervention's effectiveness. Online repositories of EBIs use a variety of evidence rating systems. (See Design and Effect size.)

Experimental condition. In an intervention study, the experimental condition or group is the one that receives the intervention. Outcome scores from that group are compared to those of one or more other groups (control or comparison groups) that did not receive the intervention.

Experimental design. Experimental designs by definition use random assignment of the full initial group of potential participants to the experimental and control or comparison conditions. In an experimental design that is adequately implemented, researchers can claim the intervention *caused* improved outcomes and not variations in the initial characteristics of participants across conditions. Many online repositories, including What Works Clearinghouse, rate the evidence for interventions' effectiveness based on their design. Experimental studies are always considered the best. (See Design, specific designs, Control group, Comparison group, Random selection, and Random assignment.)

External validity. The external validity of an intervention study is the degree to which its findings can be generalized to other schools, districts, and student populations. EBITE's emphasis on the context in which an intervention will be implemented relates to external validity. If an intervention is not appropriate for your setting or students, it does not matter how effective it was in an intervention study. Greater external validity often is associated with less *internal* validity. (See Internal validity and Causality.)

Fidelity. Fidelity refers to how fully an intervention is implemented as intended. Many EBIs have manuals that describe in detail how the intervention is supposed to be implemented and tools for monitoring and documenting fidelity. Straying from fidelity is likely to lead to less

optimal outcomes. There can be tension between fidelity and adaptation to the cultural and contextual realities of an educational setting.

Follow-up. In many intervention studies, researchers collect data from participants one or more times after they collect post-test data at the end of the intervention. Follow-up data helps researchers determine if an intervention is associated with a lasting effect on participants. Whether or not the intervention *caused* the effect, of course, is based on the study design. (See Design and specific designs.)

Improvement Index. The improvement index is a measure of the effectiveness of an intervention used by the What Works Clearinghouse. It represents the expected change in the average outcome score of those students who *did not* get the intervention if they *had* gotten the intervention. It is basically the difference between post-test scores of those who received the intervention and those who didn't. The unit of change the improvement index is the percentile rank of students on the outcome of interest. The index is a standardized value but can be converted into the point difference on the outcome between students in the different conditions. (See: https://ies.ed.gov/ncee/wwc/Glossary/improvement%20index, Effect size and Percentile rank.)

Intent-to-treat (ITT). Intent-to-treat studies include outcome data from participants who were *supposed* to receive the treatment but didn't. For example, students in a school may all have been assigned to receive one of two interventions. Some students didn't like the intervention, declined to take part, or dropped out because it wasn't helping them. To best capture the real-life effects of the intervention, data from all the students assigned to each condition should be included. The possible lack of uptake needs to be taken into account when a school or district is selecting interventions. Online repositories with detailed descriptions of evidence for interventions might refer to ITT analyses. Positive effects are harder to demonstrate with ITT analyses because some students included in the analyses didn't receive the intervention.

Internal validity. The internal validity of an intervention study is the degree to which its design and implementation allow for claims of causality. Therefore, well-executed experimental designs (with random assignment to conditions) have the most internal validity. Unfortunately, maximizing a study's internal validity often reduces its external validity because only a narrow group of students in a carefully selected setting is targeted. (See Causality, Experimental design, Designs, and External validity.)

Logic model: A logic model is one important tool for planning and monitoring the implementation of an intervention. It details the sequence, resources, activities, intended outputs, and expected outcomes of the intervention. A logic model provides a graphical overview of the intervention process. (See EBITE logic model lesson and resources.)

Longitudinal design: Outcome data in longitudinal study designs are collected at more than two data points allowing for an examination of when and how much change happens over the

course of the study and/or at one or more time points after the study ends. There are special analysis methods for evaluating longitudinal intervention effects. (See Design and specific designs.)

Mean: The mean or average of an outcome score for a group of students or study participants is the sum of the scores divided by the number of individuals. Comparison of the means of participants in different conditions is the most common way of evaluating interventions and calculating effect sizes. Reporting school-level mean scores is a common way of evaluating schools' performance. However, mean scores say little about how many students within a group or school have improved or desirable scores. Therefore, means are not useful for informing schools which students may or may not need intervention. Some scholars recommend looking at percentages of students with different ranges of scores or the mode of a set of outcome scores. It is also important to disaggregate means and percentages by demographics. (See Disaggregating data, Mean, Median, Mode, and Standard deviation.)

Median: The median score in a list of outcome scores arranged from lowest-to-highest is the point at which half of the scores are above and half are below. Median scores can be informative in evaluating students' performance or comparing intervention groups, but are not commonly reported in online repositories. (See Mean, Mode, and Standard deviation.)

Meta-analysis. A meta-analysis is a systematic, quantitative study of studies. In intervention research, meta-analyses are used to synthesize the findings of multiple studies of the same or similar interventions. They transform quantitative findings across the studies into comparable units (effect sizes) so they can be summarized, often graphically. Meta-analyses usually report the significance and magnitude of effects of multiple outcomes, because most interventions target more than one narrow outcome. When well-implemented, meta-analyses provide rigorous evidence of the effectiveness or lack of effectiveness of interventions. It should be noted, that they often reveal a range of results, suggesting differences in implementation, context, and populations.

Mode: The mode of a set of outcomes scores is the value that occurs the most often. The mode of outcome scores can be useful in cases where outcomes can be categorized, e.g., into "1=got worse," "2=stayed the same," or "3=improved." If the most common category among students in the experimental group is "3=improved" compared to "2=stayed the same" for the control group, the intervention may have been effective. (See Mean, Median, and Standard deviation.)

Multi-tiered systems of support (MTSS). Different names have been used for multi-tiered systems of support (e.g., Response to Intervention, RTI; Positive Behavioral Interventions and Supports, PBIS), but they all refer to three levels of intervention/prevention. Universal, or tier 1 strategies, occur at the school level to benefit all students and prevent the need for more intensive intervention among the majority of students. Selective, or tier 2 strategies, are interventions for students who need more targeted intervention in addition to tier 1 strategies. They are usually implemented at the group level. Indicated, or tier 3 interventions, are for

students who need supports beyond tiers 1 and 3. Special education services are tier 3 interventions, but not all tier 3 interventions are special education services.

Non- or pre-experimental design. Non-experimental designs do not have characteristics supporting claims of causality. Cross-sectional studies, pretest/post-test studies with only one group, and post-test-only studies are examples of non-experimental designs. Non-experimental studies do not provide evidence of intervention effects, but they may provide information valuable for designing interventions at ESSA's levels 3 and 4.

P value: The *p* value of a statistical finding in intervention research indicates the probability that the conclusion of a positive intervention effect is wrong. The estimate is a possible but improbable value given the true population value. Most intervention researchers use a *p* value of .05, meaning they are willing to accept a 5% chance that their conclusion of effectiveness is wrong. The smaller the *p* value of an estimate, the less likely it is that an estimate is wrong. Because there is always a statistical chance that conclusions in one study are inaccurate, it is important for interventions to replicated. Online repositories will give higher ratings to interventions with consistent positive effects across multiple studies. (See Bias, Replication, and Standard error.)

Percentile rank. The percentile rank of a student is based on scores from a normed measure—that is, scores on an outcome measure obtained from an external group of students. In intervention research, it indicates the how well students in a study condition performed relative to the external group of students used to norm the outcome measure. A student with a percentile rank of 85 performed as well as or better than 85 percent of students in the normed group. The average percentile rank of students in an intervention group is compared to the average in a control or comparison group to determine how much change in performance the intervention was associated with. That change score can be converted into an ES that can be compared across studies. (See Effect size and Improvement index.)

Post-test: Post-test data are data collected at the end of an intervention. Post-test scores can be compared to pretest scores and to post-test scores of other groups. (See Follow-up, Pretest and Pretest/Post-test design.)

Power (of a study): The power of an intervention study refers to its ability to determine that an intervention had an effect. Sample size and other statistical considerations determine the effect size a study may be able to detect. It is harder to detect a small effect (e.g., .20); and easier to detect a big effect (e.g., .60). Sample size is an important factor in power; with larger samples, studies can detect smaller intervention effects. Researchers want to avoid the situation where an intervention actually has a beneficial effect, but they don't see it because the number of study participants was too small.

Pretest: Pretest data are collected at the beginning of an intervention. They can be compared across experimental and control/comparison groups to see if the groups are equivalent before receiving the intervention. They are also compared to post-test data to see if outcomes changed between the beginning and end of an intervention. (See Post-test and Pretest/Post-test design.)

Pretest/Post-test design: Studies with pretest/post-test designs have data collection before and immediately after an intervention is implemented. Researchers are hoping that post-test scores (usually means) are better than pretest scores. The design can be used with one group or more than one group. To claim causality, two equivalent groups (obtained through random assignment) either receive or don't receive the intervention and have their post-test scores compared. Scores capturing change from the beginning to the end of the intervention can also be compared. (See Causality, Design, Equivalent groups, Post-test, Pretest, Follow-up.)

Quasi-experimental design: In contrast to experimental designs, quasi-experimental designs do not use random assignment of recruited participants to the intervention and comparison groups. Participants end up in conditions based on some other procedure or situation, such as first-come, first served; matching (finding comparison groups that are similar to the intervention group); willingness of a setting to take part in an intervention; parents who give consent to take part in an intervention; etc. Schools often decline to allow random assignment studies because it means denying potentially beneficial services to some students. Depending on the quality of the comparison group used, a quasi-experimental design could be considered almost as rigorous as an experimental design. (See Comparison group, Control group, Experimental design, and Random assignment.)

Random assignment: Random *assignment* refers to how study participants are placed in either the intervention group or control group. It means that every potential participant has the same chance of ending up in either condition. Random assignment can occur during recruitment of participants, for example, if students who are referred for services are alternately assigned to condition; or after the recruitment of a pool of study participants. In the latter case, assignment can be based on the flip of a coin or the use of a random number generator. With random assignment, researchers have confidence that the two conditions are equivalent at the beginning of the intervention. Then, differences in outcomes scores can be attributed to the intervention. (See Causality, Comparison group, Control group, Design, and Internal validity.)

Random sampling: Random *sampling* refers to how potential participants are recruited to take part in a study and before they are assigned to conditions. If researchers don't have enough money to provide an intervention to every classroom in a school, or every school in a district, for example, they might use a random process to select schools and/or students. Random selection is desirable because, if implemented well, it allows researchers to claim their findings from a subset of potential participants apply to the entire set—there is no reason to believe that the excluded units were different from the included units. (Compare to Random assignment.)

Regression. Regression analysis is one type of statistical analysis that can be used to examine the effects of an intervention on an outcome. (See ANOVA, ANCOVA, Mean, and *t*-test.) Instead of comparing group means, regression analysis looks at how much being in one condition versus another predicts outcome scores. Regression analyses can easily identify intervention effects over and above the potential effects of other variables, such as gender, age, and pretest scores. It can also be used to see if an intervention works better for participants with different characteristics. Evidence for ESSA levels 3 and 4 are often based on results of non-intervention studies using regression analyses.

Replication/Replicability. Positive evidence for an intervention is considered much stronger when more than one study has been conducted and found beneficial effects. Ideally, replication studies are separate studies conducted by researchers who were not part of the team that developed the intervention. Many interventions have not been replicated, especially by independent researchers (e.g., non-developers). Some online EBI sites report whether replications have been conducted and use them as a rating criterion. However, it is not uncommon to find interventions in the repositories that have not been replicated or that were replicated but replication findings were less positive than those found in the original study. Inconsistent findings reduce confidence in the effectiveness of an intervention. (See *P* value.)

Reproduction/Reproducibility. Reproducibility refers to whether or not the findings of an intervention study can be reproduced in a re-analysis of the study's data by separate researchers. When re-analysis of study data leads to the same results as the original study, researchers and practitioners can have more confidence in the reliability of the original research findings.

Single-subject or **Single-case design.** Single subject designs have been used extensively in educational research. In this type of study, the outcome(s) of one student or one group of students is measured multiple times during a baseline period (A) and plotted on a graph. The baseline trend is measured until a clear pattern of scores is obtained. Then an intervention phase is begun (B). The outcome continues to be measured regularly and graphed. The individual or group is its own control group because the trend in outcome scores in the intervention phase is compared to the baseline trend. Simple statistical procedures are used to see if outcome scores during baseline and intervention phase are statistically significant. The most rigorous single-subject designs use multiple A and B phases to see if outcome trends are consistently different for baseline (no intervention) and intervention phases. Using a single-subject design to simultaneously study AB trends in outcomes for multiple students or groups is also a rigorous design—the study of each student or group can be considered a replication of the intervention study. The What Works Clearinghouse considers well-implemented single-subject designs as experimental studies.

Standard deviation (*SD*). The standard deviation of a set of scores gives information about the spread of scores around the mean. Technically, 67% of scores in a set fall within + or – one SD of the mean. The size of the *SD* of outcome scores for two experimental groups affects whether

their mean differences are statistically significant. The wider the *SD* of each group's outcome scores, the harder it will be to determine their means are statistically significantly different. Some online EBI repositories will report *SDs* from intervention studies, and some scholars say they are important numbers to consider when evaluating effects of an intervention because they may demonstrate the intervention had a wide range of effects. (See Mean, Median, and Mode.)

Standard error (SE). The standard error of an estimated intervention effect indicates how precise or accurate the estimate is; that is, how much error there is in the scores. An estimate is an average score for all participants in the study. The farther each participant's score is from that average estimated effect, the less precise the estimate is across the whole sample and the more error it contains. The *SE* is part of the calculation of statistical significance—the larger the *SE*, the less likely the estimate effect is to be statistically significant. (See *P* value.)

Statistical significance. See *P* value.

Treatment-as-usual (TAU). In some intervention studies, the effects of an intervention are compared to the effects of treatment-as-usual—that is, the services students would get in the absence of the new intervention. For example, students in an experimental group may receive a new reading curriculum, while students in a control or comparison group continue to receive last year's curriculum. (There may also be a group who receives no intervention, but that is not likely to be acceptable to school staff.) At the end of the school year, group reading scores can be compared to see if the intervention led to significantly better reading scores. Some intervention reports in online repositories may refer to TAU comparisons. In general, if the usual treatment had any positive effect at all, it will be more difficult to demonstrate a significant positive effect of the new intervention.

***T*-test.** T-tests compare two group means to see if they are statistically significantly different. There are only two variables in an intervention study *t*-test: the variable with information on intervention group (with two categories) and the outcome variable. T-tests are commonly used for intervention studies. As with any means comparison, they summarize results for the students in each condition. They do not give specific information on how many students benefited from the intervention or characteristics of those who benefited or did not benefit.