Threats to External and Internal Validity

The results of a USAID evaluation are only useful to the extent that they can be accurately and confidently interpreted and applied to programs. **Validity**, which is derived from the Latin word *validus,* meaning "strong," refers to the degree to which correct inferences can be made from the results of an evaluation. The idea of validity in an evaluation involves two main concepts. First, USAID and other evaluation users need to have confidence that the evaluation findings are a function of the conditions observed and measured and not due to other factors that were not addressed in the evaluation. Such confidence reflects the **internal validity** of an evaluation. Second, USAID usually wants to use the results of an evaluation to make a claim not just about the participants being observed but also about a larger group of current or future program participants. The ability to make such claims, or generalizations, depends on the **external validity** of the evaluation.

There are often important tradeoffs between the costs of evaluations and accuracy. USAID decision-makers need to know when a high degree of internal and external validity is paramount (for example with high stakes impact evaluations aimed at major resource allocation/scale-up decisions) and cases where internal and external validity should be considered and managed less formally within the time and budget constraints that are typical for most performance evaluations. Being an informed consumer of evaluations involves understanding some of the common threats to validity and how they affect the accuracy and generalizabilty of specific evaluations. USAID Evaluation Specialists are well placed to understand threats to validity and ways to mitigate them as part of the evaluation design.

PART I: External Validity--*The ability of a given design to produce results that can be held to be true for other cases, such as different people, places or times.*

Threats to the external validity of evaluation findings may be related to the **population**, that is, the extent to which a sample is representative (or not representative) of the program population from which it was selected, or to the **ecology**, that is, the extent to which characteristics of the setting or context of the evaluation are representative (or not representative) of the setting and context to which the results are to be generalized. The following section describes some of the common threats to external validity and ways to mitigate them.

In general the threats to external validity can be minimized when the evaluators take steps to ensure that the people, the setting, and the context are representative of the population, setting and context to which the findings will be generalized.

	Threats to External Validity	***Ways to Mitigate Threats***
Effects of Testing	Refers to the fact that a test (for example, a pretest) may affect the responses or the performance of the participants in an evaluation	The use of evaluation designs that do not include pretests can help eliminate this potential threat. Evaluation designs such

	or research study. If this happens, it means that the performance of the people being studied may be different from what it might have been if they had not been pre-tested. Therefore, the results may not be generalizable to situations where pre-testing will not occur.	as the Solomon four-group design are especially useful in determining the extent to which pre-testing may have influenced the results of a study.
Multiple Treatment interference	A situation in which participants in an evaluation receive more than one treatment. In such a case, the effects of the multiple treatments may interact. For example, suppose an evaluation is examining the productivity of small farmers who are receiving assistance from several aid agencies. Those other "treatments" may have some impact on farmer productivity. The results of this evaluation can be validly generalized only to similar situations.	When there is reason to believe that there will be interference of multiple treatments, the evaluators should try to choose a design in which only one treatment is assigned to each subject. In an impact evaluation, the researcher should try to control and/or measure the effects of all relevant treatments and incorporate them into a multiple-treatment design. In a performance evaluation it may only be practical to recognize and describe that several treatments are in play and are likely contributing to the results.
Selection-treatment interference	The possibility that some characteristic of the participants selected for the evaluation interacts with some aspect of the treatment. Examples of such characteristics could include prior experiences, learning, personality factors, or any traits that might interact with the effect of the treatment. For the results to be validly generalized to a larger population, that population must possess the same traits, characteristics, experiences, and so on as the sample.	This threat is similar to the internal validity threat of selection, and the remedy is also similar. Random selection and assignment of participants can minimize much of the threat to external validity due to selection-treatment interaction. When random selection or random assignment is not practical, statistical techniques such as analysis of covariance, used in conjunction with a careful quasi-experimental design, can take into account differences due to measurable attributes of the individual, thus minimizing selection-treatment interaction as a threat. The key is ensuring that

		the study sample is representative of the target population.
Hawthorne and other Experimental Effects	A situation where participants become aware that they are involved in a study, and, as a result of that awareness, their response or performance is different from what it would have been otherwise. The effect on performance may be due to the "newness" of the experimental treatment (sometimes called the "novelty effect"), to the belief on the part of participants that they are receiving some "special" treatment (sometimes referred to as the "Hawthorne effect";), or to the participant's belief in the effectiveness of the treatment (sometimes called the "placebo effect").	This threat is similar to the threat of 'Effects of Testing' and can be addressed by similar methods. While it is important to be transparent about participation in research, by minimizing the obtrusiveness of the study, we can minimize the threat of the Hawthorne effect.
Experimenter Effects	Refers to the possibility that an evaluator may sometimes unintentionally influence the performance of participants in a evaluation or research study. Rosenthal (1996, p. 40) classified these effects as "passive" (e.g., the gender, race, or personal attributes of the researcher or observer affect participant's performance) or "active" (e.g., the expectations of the evaluator or observer are communicated to the participant in a manner that affects performance.)	The most effective way to minimize the reactive effects of various experimental arrangements is to have a control group (i.e., a group that receives no treatment whatsoever) and a "placebo" group (i.e., one that receives a "placebo" or non-experimental treatment.) In some settings, it is often impossible to have "true" control groups, but we can usually arrange for a "placebo" group. An example would be a case in which an experimental group receives the "new" method of instruction and the "placebo" group would receive the traditional method of instruction.
Specificity of Variables	Is concerned with the extent to which the variables in study are adequately described and operationally defined. Variables	Careful definition of variables is the key to minimizing this threat to external validity. In order to ensure generalizability, the

	can be defined too specifically. For example, if an evaluator defines poverty as annual per capita income below $1000, any results may not be validly generalizable for other definitions of poverty. Minimally, all variables must be described in sufficient detail to allow another evaluator to replicate the study. In addition, the description and definition of variables must employ measurement instruments or observational devices that are themselves reliable and valid. To the extent that the variables included in an evaluation are not adequately described and carefully defined, the ability to validly generalize the results of the study is threatened.	evaluator must operationally define variables in a way that is meaningful in settings beyond that in which the study is being conducted. The use of widely agreed upon definitions or multiple competing definitions should be considered.

PART II: Internal Validity--*The ability of a given evaluation design to rule out other explanations for the observed results.*

	Threats to Internal Validity	***Ways to Mitigate Threats***
History Effect	When there is a possibility that events that occurred during the course of the intervention or between repeated measures, that are not part of the intervention, may have influenced the outcomes. For a group of individuals a historical threat to internal validity must identify an event that simultaneously affects most of the individuals enough to appreciably change the measured trait.	The use of a control or comparison group, selected from the same population as the intervention group (s) and which experiences the same concurrent history as the intervention group(s), can help eliminate most of the effects of history. Also, the shorter the duration of an evaluation, the less likely history will be a threat.
Maturation Effect	This occurs when results are caused by aging or development. Changes that naturally occur as a result of the passage of time include growing older, getting smarter	The effects of maturation, like the effects of history, can be minimized by the use of a control or comparison group, selected from the same population as the intervention group(s). Also, like

	and gaining experience. Maturation may be conceived as occurring in two forms: short or long term. **Short-term maturation** is demonstrated by fatigue and learning. **Long-term maturation** deals with psychophysical development, cultural changes, and environmental changes that can affect psychological constructs.	the effects of history, the effects of maturation tend to be minimized in programs or evaluations of short duration.
Repeated Testing Effect	This occurs when subjects are given the same test before and after the intervention or on multiple occasions. The subjects may learn how to respond to the questions, marring the validity of the results.	The use of an evaluation design that does not include a pretest can eliminate testing as a potential threat to internal validity. If baseline or pretreatment data are needed, and they are typically recommended, the use of unobtrusive measures (data collection techniques about which the participant is unaware) may minimize the effects of testing. Evaluators may also use different equivalent forms of a test for pre-testing and post-testing such as those used in this course.
Selection Bias	Selection bias is created when a group of people self-select into a program, and they are compared against another group who did not participate. Distortion of evidence or data about the results of a program intervention can occur due to <u>systematic differences</u> in the characteristics of the subset of the population receiving the intervention and the comparison group.	Random selection and random assignment of subjects minimize selection as a threat to internal validity. If random selection and assignment are not possible, the use of certain statistical techniques, such as propensity score matching, can adjust for group differences and thereby minimize selection as a threat.
Attrition (or Mortality) Effect	Mortality effect is the effect of drop-outs on the measurement of results of an intervention. Just as selection	Choosing large groups (or over sampling where high levels of mortality are expected) and ensuring that they are

	can be a source of bias, differential dropout rate (e.g. the most motivated participants leave the study area to seek new job opportunities) among participants can harm the credibility of the results.	representative of the population from which they were selected can minimize mortality threats. The use of follow-up procedures with a portion of those who leave the study or who were initially unavailable can further minimize mortality as a threat.
Regression to the Mean Effect	A statistical phenomenon that shows a natural tendency for individuals who score either very high or very low to move closer to the middle when retested. When high or low scorers on a pre-test are selectively sampled, we expect slight decreases or increases in their scores naturally. This effect is known as regression to the mean or the average.	Including a control group of similar individuals can help identify and control for regression to the mean. When this is not possible, more extreme samples, either at the top or bottom of the distribution of initial scores, will exhibit larger regression to the mean, so including a wider range of the sample reduces this threat.
Instrumentation Effect	This occurs if the reliability of the instrument changes. Changes can be a result of measurement scale or standards changes or recalibration of instruments of measurement (e.g. recalibrating a health clinic's weight scale during a nutrition intervention). This causes a lack of reliability of the instrument used to measure indicators.	Careful specification and control of the measurement procedures can eliminate most instrumentation threats. Standardized instruments, administration or data collection procedures, and the training of observers/enumerators are among the procedures that help control the instrumentation threat.

Sources:

Imas, Linda G. Morra and Ray C. Rist, *The Road to Results: Designing and Conducting Effective Development Evaluations*. The World Bank. Washington D.C.: 2009

Bieger, George, Pennsylvania Indiana University. On-line Course on Research Methods. Analysis of Research in Literacy

Martella, R. and Ronald Nelson (1999). *Research Methods, Learning to Become a Critical Consumer*. New York: Allyn and Bacon.