Evaluation Terms

Created 3/25/11

Evaluation terminology can be confusing. This resource provides commonly accepted definitions of key evaluation terms. This is not an exhaustive list. For more information about evaluation, go to Research Methods Knowledge Base Evaluation Research or Selected Evaluation Terms from Harvard Family Research Project.

Types of Evaluation

Two common definitions of evaluation include the following:

1. "Evaluation is the systematic assessment of the worth or merit of some object." An "object" may include a sexual violence primary prevention program or curriculum.
2. "Evaluation is the systematic acquisition and assessment of information to provide useful feedback about some object."

There are two major types of evaluation (formative and summative evaluations). The approach used depends on the "object" being evaluated and the purpose of the evaluation.

Formative evaluations strengthen or improve what is being evaluated. They examine the delivery of a program and quality of implementation as well as assess the organizational context.[1] These evaluations are often conducted during program implementation to provide information that will improve the program. Findings point to aspects of program implementation that may be improved for better results (i.e., how a primary prevention program is administered or how health educators are trained).[2]

Summative evaluations examine the effects or outcomes of a program, describe what happens following delivery of the program, and assess whether the program may have contributed to the outcome.[3] These evaluations are often conducted during or at the end of a program's implementation and may be used to determine whether a program's intended outcomes were achieved.[4] An example is a questionnaire or post-test administered after a sexual violence primary prevention program is implemented. The data would be analyzed to determine the effects of the program on outcomes or behaviors of interest.

---

[1] Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition. Retrieved from Internet on March 25, 2011 from http://www.socialresearchmethods.net/kb/index.php.

[2] Little, P. March 2002. Harvard Family Research Project, Selected Evaluation Terms. Harvard Family Research Project. Retrieved from Internet on March 27, 2011 from http://www.hfrp.org/evaluation/publications-resources/selected-evaluation-terms).

[3] Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition. Retrieved from Internet on March 25, 2011 from http://www.socialresearchmethods.net/kb/index.php.

[4] Little, P. March 2002. Harvard Family Research Project, Selected Evaluation Terms. Harvard Family Research Project. Retrieved from Internet on March 27, 2011 from http://www.hfrp.org/evaluation/publications-resources/selected-evaluation-terms).

General Terms Used in Evaluation

The following definitions were adapted from existing resources on program evaluation. [5,6,7]

Attrition is the loss of participants during the period of time when a program is being implemented (i.e., drop outs). An example of attrition would be if students did not complete the entire sexual violence primary prevention program or did not attend all sessions provided in the curriculum).

Bias is a systematic error that can underestimate or overestimate findings in some way. For example, if a sexual violence prevention program administers a post-test to only those who have successfully completed the program, results will be *biased* towards the experiences of those who successfully completed participants and would not reflect those of participants who dropped out.

Code is to translate a set of data into quantitative values or qualitative categories. For example, you may assign numeric codes to responses on a survey, such as yes=1 and no=0.

Correlation is a single number that describes the degree of relationship between two variables. For example, you may calculate the correlation between age of the participant and rape supportive attitudes.

Dependent variable is the measure (e.g., behavior) that is assumed to vary as a result of some influence (e.g., primary prevention program). For example, in bystander education program, the *dependent variable* is the likelihood that a participant will intervene when witnessing an act of sexual violence. (Please see *independent variable.*)

Design is the plan for how an evaluation will be conducted (i.e., questions to be addressed, data collection, and analysis). For example, you would design a plan on how you would evaluate a sexual violence primary prevention curriculum for middle school students.

Generalizability **or *external validity*** is the extent to which information about a program designed for one group may apply to other settings, people, and time. For example, you would determine if evaluation findings/outcomes from a school-based program in Arizona applied to students in California.

Independent variable is a program or intervention (e.g. primary prevention program) that may influence a specific behavior. For example, participation in a bystander education program is the *independent variable* that ideally influences willingness to intervene among participants. (Please see *dependent variable.*)

[5] Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition. Retrieved from Internet on March 25, 2011 from http://www.socialresearchmethods.net/kb/index.php.

[6] Glossary of Program Evaluation Terms. Western Michigan University. Evaluation Center. Reprinted from *The Program Evaluation Standards, 2nd Edition* (Sage, 1994). Retrieved from Internet on March 28, 2011 from http://ec.wmich.edu/glossary/prog-glossary.htf#Table of Contents)

[7] Glossary of Evaluation Terms. (2009). Planning and Performance Management Unit. Office of the Director of U.S. Foreign Assistance. Retrieved from the Internet on July 20, 2011 from http://pdf.usaid.gov/pdf_docs/PNADO820.pdf.

Internal Referencing Strategy (IRS) is a single group comparison method of evaluating a training to measure effectiveness. Pre- and post-tests are administered to training participants with measures relevant (planned changes) to the training as well as irrelevant measures (unplanned changes). The training is effective if changes to the relevant measures are greater than the changes to the irrelevant measures.[8] For example, participants attending rape myth training should decrease their acceptance of common rape myths. Participants will also probably reduce gender stereotyping as a result of the curriculum, however this behavior change is not the intent of the training. If the decrease in rape myth acceptance is greater than the decrease in gender stereotyping, the training was successful.

Levels of measurement. The type of statistical operations employed depend on how variables are measured.[9]

- **Nominal**. Numbers or symbols assigned to a set of categories for the purpose of naming, labeling, or classifying the observations (i.e., Caucasian = 1, African American = 2, Latino = 3, Native American = 4). [10]

- **Ordinal** variables are ranked from low to high. Likert scales are examples of ordinal variables (i.e., strongly agree = 5 to strongly disagree = 1).[11]

- **Interval-Ratio**. Variables where measurements for all cases are expressed in the same units. Variables with a natural zero point, such as height and weight, are called ratio variables (i.e., age, income, SAT scores). [12]

Measurement is the process of observing and recording the observations collected. For example, measurements may include survey questions about bystander behaviors.

Population is the group you want to generalize the findings to. For example, if you administered a primary prevention program to a class of 7th graders in Phoenix, you would want to determine if the results were generalizable to other classes of 7th graders in other geographic regions or settings. The population in this example would be classes of 7th graders.

Random assignment is how you assign individuals in your sample to different groups. Each person has an equal chance of being assigned to one of the groups (similar to flipping a coin). Groups may include a prevention group (i.e., the group who received the primary prevention program) and control group (i.e., the group who did not receive the primary prevention program) in your program.

Random selection is how you draw the sample of people for your program from the target population. You chose a number of individuals from a population so that all individuals in the

[8] Haccoun, R. R & Hamtiaux, T. (1994). Optimizing knowledge tests for inferring learning acquisition levels in single group training evaluation designs: The internal referencing strategy. *Personnel Psychology.* 47(3):593-604.
[9] Hussaini, K. (2011). *Data entry and analysis.* [PowerPoint slides]. Office of Assessment & Evaluation (BWCH).
[10] Hussaini, K. (2011). *Data entry and analysis.* [PowerPoint slides]. Office of Assessment & Evaluation (BWCH).
.
[11] Hussaini, K. (2011). *Data entry and analysis.* [PowerPoint slides]. Office of Assessment & Evaluation (BWCH).
[12] Hussaini, K. (2011). *Data entry and analysis.* [PowerPoint slides]. Office of Assessment & Evaluation (BWCH).

population have the same chance of being selected and would be a representative of the target population.

Random error may be caused by factors that randomly affect measurement of the variable in the sample. For example, a student's mood may affect the way they answer questions in a sexual violence primary prevention program survey on that particular day.

Response is a specific measurement value that a sampling unit (i.e., person) provides. For example, a student provides a response to a question on a sexual violence primary prevention survey.

Sample is the group of people selected to participate in a program and may not include all people who are actually *in* the program. For example, some people may drop out over the course of a multi-session program. The group that actually completes the entire program is a *subsample* of the sample and does not include non-respondents or dropouts.

Sampling is the process of selecting units (e.g., students, organizations) from a population of interest. Usually, you want to generalize your results back to the population from which they were chosen.

Surveys are measurement procedures that involve asking participants questions. They may consist of a short paper-and-pencil feedback form or an individual interview.

Systematic error may be caused by factors that affect measurement of the variable across the group of participants (not just one individual). For example, if there was loud traffic outside of the classroom, students' scores on a pre-test may be affected by the distraction.

Test-retest reliability is the degree to which you get consistent results when a test is administered twice to the same group of participants. Strong test-retest reliability is achieved when you administer a sexual violence primary prevention survey to the same group of students twice and you obtain the same results.

Variable is anything that can take on values. For example, age is a variable because age can be represented by different values for different people or for the same person at different times. Gender is also a variable and may consist of three values: male, female, and transgender. Other examples include income level and counties which people live in. Household income level may be measured as a variable using categories such as below $15,000 a year, between $15,000-$24,000, between $25,000-$34,999, etc.

Validity & Reliability[13]

Validity is "the agreement between a test score or measure and the quality it is believed to measure."[14] It measures the gap between what a test *actually* measures and what it is *intended* to measure. For example, the Illinois Rape Myth Acceptance Scale (IRMA) is a widely used tool to assess participants' level of acceptance of common myths about sexual assault. Its validity was measured by conducting a series of studies comparing the relationship of the IRMA to other studies and theories related rape acceptance variables.[15]

External validity or *generalizability* is the extent to which the conclusions would hold true for other persons in other places and at other times. For example, a social norms intervention to prevent sexual aggression amongst college men may not be externally valid to high school students, but a program using a Peer Education Model can be applicable to a variety of populations.

Internal validity is the degree to which the conclusions made in the evaluation are supported by the evidence collected.[16] For example, bystander behaviors are associated with participation in a sexual violence primary prevention program and not information received in the public media.

Reliability is the consistency or repeatability of measures. A measure is reliable if we get the same result when it is administered multiple times. The Illinois Rape Myth Acceptance Scale (IRMA) was also proven to be a reliable tool by ensuring that there was minimal potential for bias, items were clearly understood, and that colloquial phrases were kept at a minimum.[17] (See *validity*)

Qualitative vs. Quantitative[18]

Qualitative data has different forms. It includes virtually any information that may be captured that is not numerical in nature, such as open-ended interviews and observations. For example, students who participated in a primary prevention program may be asked to describe in their own words how the program changed their attitudes and perceptions of survivors of sexual violence.

Quantitative data may be assigned numerical values that may be used in statistical analyses. For example, the biannual Youth Behavior Risk Surveillance Survey gathers *quantitative data* on the

[13] Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition. Retrieved from Internet on March 25, 2011 from http://www.socialresearchmethods.net/kb/index.php.

[14] Kaplan. R. M., & Saccuzzo, D. P. (2001). *Psychological Testing: Principle, Applications and Issues (5th Edition)*, Belmont, CA: Wadsworth. Retrieved from Internet on may 14, 2011 from http://changingminds.org/disciplines/hr/selection/validity.htm.

[15] Payne, D. L., Lonsway, K. A., & Fitzgerald, L. F. (1999). Rape myth acceptance: Exploration of its structure and its measurement using the Illinois Rape Myth Acceptance Scale. *Journal of Research in Personality, 33,* 27-68.

[16] Glossary of Evaluation Terms. (2009). Planning and Performance Management Unit. Office of the Director of U.S. Foreign Assistance. Retrieved from the Internet on July 20, 2011 from http://pdf.usaid.gov/pdf_docs/PNADO820.pdf.

[17] Payne, D. L., Lonsway, K. A., & Fitzgerald, L. F. (1999). Rape myth acceptance: Exploration of its structure and its measurement using the Illinois Rape Myth Acceptance Scale. *Journal of Research in Personality, 33,* 27-68.

[18] Little, P. March 2002. Harvard Family Research Project, Selected Evaluation Terms. Harvard Family Research Project. Retrieved from Internet on March 27, 2011 from http://www.hfrp.org/evaluation/publications-resources/selected-evaluation-terms).

percentage of youth who reported having ever been physically forced to have sexual intercourse. What is the difference between quantitative and qualitative data?

	Uses	Sample Methods
Quantitative Data	used to compare outcomes associated with an intervention	tests/assessments secondary source/data review (i.e., school records, medical records) surveys/questionnaires
Qualitative Data	used to understand how a program operates and how participants experience the program	document review individual interviews focus groups observation