Stage 5: Gathering Data and Assessing Results

What Does Evaluation Measure?
Methods of Data Collection
- Surveys
- Interviews
- Observations
- Records
- Meetings

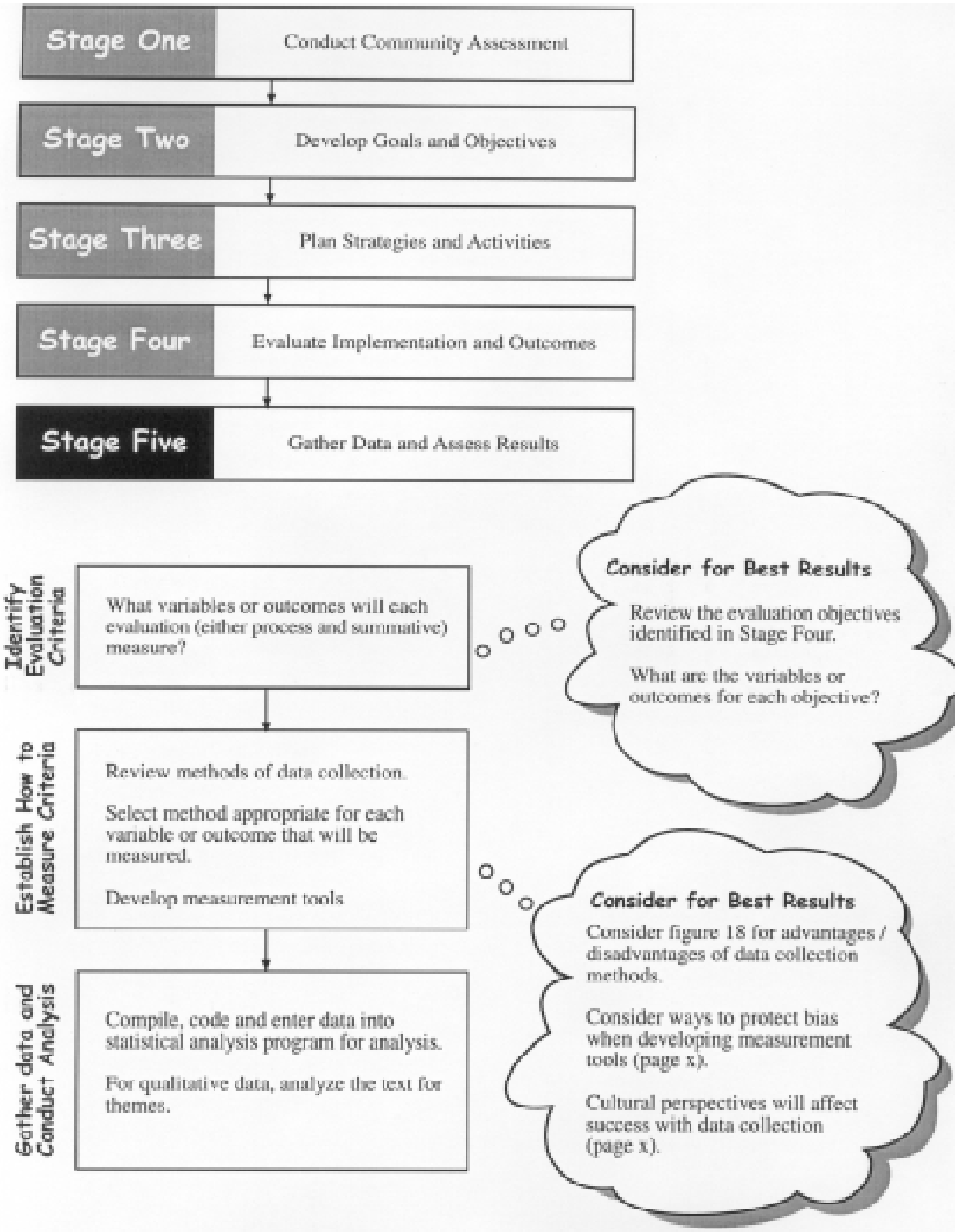
Quality of data collection
- Reliability
- Validity
- Cultural appropriateness

Data Analysis
- Coding
- Quality control

Types of Analyses
- T-tests
- Univariate analysis
- Bivariate analysis
- Multivariate analyses

Figure 14: Selected Evaluation Criteria
Figure 15: Methods for Collecting Data

Toolkit
References
Workform for Ways to Measure Process
Workform for Ways to Measure Outcomes


Stage One
Conduct Community Assessment
Stage Two
Develop Goals and Objectives
Stage Three
Plan Strategies and Activities
Stage Four
Evaluate Implementation and Outcomes
Stage Five
Gather Data and Assess Results
Identify Evaluation Criteria
What variables or outcomes will each evaluation (either process and summative) measure?
Establish How to Measure Criteria
Review methods of data collection.
Select method appropriate for each variable or outcome that will be measured.
Develop measurement tools
Gather data and Conduct Analysis
Compile, code and enter data into statistical analysis program for analysis.
For qualitative data, analyze the text for themes.
Consider for Best Results
Review the evaluation objectives identified in Stage Four.
What are the variables or outcomes for each objective?
Consider for Best Results
Consider figure 18 for advantages / disadvantages of data collection methods.
Consider ways to protect bias when developing measurement tools (page x).
Cultural perspectives will affect success with data collection (page x).

Thus far, Stages 1-4 have described program planning considerations for development and implementation of outreach activities and for evaluating what is accomplished and what can be improved. Assessment of actual implementation and outcomes, called process and summative evaluation, provides accountability and helps inform program decisions or improvements. Stage 4 addressed several considerations for planning how process and summative evaluation will be conducted, including:

- Determining evaluation objectives
- Determining more specific priorities for what should be discovered, tested, or verified
- Determining types of data to collect, when, and from whom

In Stage 5, evaluation planning continues by considering what evidence will be measured or observed and how to best measure or observe it. This chapter will address methods of collecting data and analyzing results.

What Does Evaluation Measure?

The basic question answered by measurement and analysis is how data collected from the program compares with program evaluation criteria. *Program evaluation criteria* are what determine evaluation objectives and answers to questions posed by you and your stakeholders.

Thus, criteria that evaluation might measure, depending on what you want from the evaluation (as discussed in Stage 4), include:

- Outreach objectives – if carefully constructed, as seen in Stage 2, each objective includes specific indicators and criteria;
- Characteristics of the outreach process considered important for reaching success (addressed in process evaluation);
- Information about implementation that is important for program replication (addressed in process evaluation);
- Assumptions about cause and effect of strategies – relationship between independent and dependent variables ;
- Outcomes not already measured in outreach objectives .

In planning for data collection, think broadly about which evaluation criteria correspond to what you and your stakeholders want to find out.

Figure 14: **Indicators of Selected Outreach Objectives**

Type of Indicator	Example Means of Obtaining Data
Awareness	• Written instruments (e.g. true-false items, completion items) • Proxy measure (e.g. number of pamphlets picked up)
Knowledge	• Written/oral test (e.g. completion items, multiple-choice items, true-false items)
Attitudes	• Written instrument (e.g. Likert scale, cumulative scale, value scale, forced choice)
Behavior	• Self-report written instrument (e.g., completion, short-answer essay, multiple-choice, true-false) • Observation (obtrusive and unobtrusive) • Proxy measures (e.g. number of people who accessed a website, number of requests received for materials)
Skills	• Observation (obtrusive and unobtrusive) • Skills test (e.g. able to retrieve specific type of clinical research)

The next section describes the instruments and tools for various methods of data collection. To help with decision making about what criteria are measured and what methods will be used, complete the Workforms provided in Stage 5 Tool Kit. For completed workform samples, please refer to Appendices L and N.

Methods of Data Collection

(See McKenzie, 1997 (1) for a thorough description of the data collection methods covered briefly in this section.)

Written questionnaires, telephone interviews, and face-to-face interviews are methods of collecting data from *respondents*. Respondents are the individuals who supply this information, so the measures are called *self-report*. Self-reported results are always influenced by the person's ability to recall accurately("When were you last on the Internet?" and report honestly("I use PubMed daily"). Offering anonymity is helpful in gaining honest answers.

Surveys are instruments that present information to a respondent in writing or pictures requiring a written response – a check, circle, word, sentence, or several sentences. Surveys can be conducted by mail, in person, by telephone, or electronically.

Survey research is one of the most common methods used in outreach evaluation, e.g.,

- For a community or audience assessment
- For pre- and and posttests in a process evaluation to determine progress or improve quality
- For followup questions asked *after* an outreach activity to determine what has happened as a result of outreach participation

Interviews are structured dialogues conducted between two (or more) persons, in which a respondent answers questions posed by an interviewer. The questions may be predetermined, but the interviewer is free to pursue interesting responses. Focus group interviews take advantage of small group dynamics (usually eight to twelve individuals). The open-ended nature of interviews or focus groups allows participants to provide answers in their own words and allows researchers to better understand issues from the perspective of the audience.

Observations require that one or more observers devote attention to the behavior of an individual or group in a natural setting. Protocols about who or what to observe, when and how long, and the method of recording the information (e.g., a questionnaire or tally sheet) can guide observers. Or, an observer may simply record an account of events that occurred within the prescribed time period, without following a guide for what to observe, for how long, etc.

Records are systematic accounts of regular occurrences consisting of such things as sign-in sheets, interlibrary loan tallies, document service requests, computer log files.

Meetings are a good source of information for the formative planning stages of a program. For example, a meeting with contacts of the targeted audience and outreach staff will be helpful for effective planning of the implementation and evaluation. The meeting structure can be flexible to avoid limiting the scope of the information gained. Possible biases may occur if those involved feel they need to give "acceptable" responses rather than discussing actual concerns.

Figure 15 summarizes some advantages and disadvantages of various data collection methods (2).

Figure 15 **Methods for Collecting Data**		
	Advantages	**Disadvantages**
Questionnaire	• Provides answers to a variety of questions • Can be answered anonymously • Allows time before responding • Can be administered to many people, at distant sites, simultaneously • Imposes uniformity by asking all respondents the same thing	• Are not as flexible as interviews • People can often express themselves better orally than in writing • Getting people to complete questionnaires can be difficult • Good questions take time to develop and test
Interview	• Can be used for non-native speakers or those who might have difficulty with the wording of written questions • Permits flexibility and allows the interviewer to pursue unanticipated lines of inquiry • Appropriate to get in-depth information for sensitive topics	• Is time consuming • Sometimes the interviewer can unduly influence the responses of the interviewee • Limits sample size
Observation	• Can be valuable if self-report measures may not be accurate • Can be seen as a report of what actually took place presented by a neutral outsider(s)	• Presence of observers may alter what takes place • Time to develop the instrument and train observers • Time to conduct sufficient number of observations • There are usually scheduling problems • Limits sample size
Records	• Often viewed as objective and therefore credible • Set down events at the time of occurrence, rather than in retrospect • Can be unobtrusive • Can have a low impact on staff time and resources if records are already kept for purposes other than the evaluation	• May give incomplete data • Examining them and extracting relevant information can be time-consuming • There may be ethical or legal constraints in examining certain records • If records are kept only for the purpose of evaluation, may be seen by staff as burdensome
Meetings	• Good for formative evaluation • Can be low cost • Permit flexibility	• Possible bias if participants feel unable to be candid

Adapted from: *How to Assess Program Implementation*, by J.A. King, L. L. Morris, and C.T. Fitz-Gibbon, 1987, Sage Publications.

Quality of Data Collection

"Quality control" criteria to guide your data collection decisions include *reliability, validity*, and *cultural appropriateness.*

Reliability is a measure of the consistency of the data collection instrument. A reliable instrument gives the same (or nearly the same) result every time. In *test-retest reliability*, the survey should produce the same results if the same person completed it twice. *Interrater reliabilty* comes into play when information is collected by different observers or raters; there should be consistency or agreement between them about the measurements. For example, two observers should give similar scores when rating the search skill competence of class participants.

Validity refers to whether the instrument accurately measures what was intended. A valid instrument increases the chance that you are measuring what you want to measure, thus ruling out other possible explanations for the results.

For example, an issue of validity might be whether you think a follow-up questionnaire can accurately measure use of PubMed for clinical decision making. Respondents may want to answer in a way that will reflect well on themselves, while not being very realistic.

To rigorously establish the validity and reliability of data collection methods gets into a technical area that may require outside assistance. For a thorough description of *instrumentation*, the technical term for selecting or developing measuring devices, readers are referred to Issac, 1995 (3). For example, Isaac describes tests for item analysis and reliability and various types of validity, including content, construct, and criterion-related validity.

However, if you are not hoping to make generalizations based on statistical validity, it is not necessary to rigorously test your data collection instruments. But, trying to be as consistent and accurate as possible is important. Reisman, et al (1994) describe how to pilot test a research instrument (4). The pilot test will answer questions such as:

- Are certain words or questions redundant or misleading?
- Are the questions culturally or otherwise appropriate for the intended respondents?
- Will the data be useable for meaningful analysis?
- Are the procedures for collecting the data clear to anyone who will do so?
- How consistent is the information obtained by the survey?
- How accurate is the information obtained by the survey?

Reisman suggests putting the instrument through a trial run with six to ten people who are similar to those likely to respond or be interviewed. Analyze the feedback from your test group to determine if questions are clear and understandable. Do people interpret the questions as intended? Are the response choices in your questions adequate and sufficient?

For example, if you know certain attitudes or behaviors of the test group subjects, are their responses consistent with their attitudes and behaviors? Select some pilot test respondents who you perceive to be uncertain about using computers to find answers to health information questions. Select a few others who you perceive to be enthusiastic about the effectiveness of using computers for health information needs. Then determine whether the questionnaire or interview responses distinguish between the two.

Cultural Appropriateness

The cultural perspectives of your targeted audience, as well as data collection strategies, should be considered in the selection process. An excellent source on

this topic is Orlandi's *Cultural Competence for Evaluators: A Guide for Alcohol and Other Drug Abuse Prevention Practitioners Working with Ethnic/Racial Communities* (5).

Members of "over-researched" ethnic minority groups, such as American Indians/Alaska Natives and African Americans, tend to be skeptical or mistrustful of the evaluation process. Their experience has been that social scientists enter their communities and collect data, but frequently fail to share their findings or take visible and beneficial action. In Hispanic communities, evaluators are viewed with suspicion as outsiders who conduct sterile research only to justify the shutdown of needed projects or services (5).

The challenge for the researcher is to build confidence in the purpose and benefits of the research results for the community. Try to involve respected community members and leaders in evaluation planning (e.g. to review a questionnaire and data collection strategy). Ask their cooperation in helping you to recruit participation. You can also directly involve members of the community in data collection efforts, such as interviews. Be sure to share your findings, if possible as early as the draft stage, for their review and comment.

Data Analysis

Once you have gathered your data from surveys, interviews, or other methods, the next steps are to conduct the analysis, draw conclusions, and prepare a report or presentation. It is important to consider how to do the analysis in the evaluation planning stage.

The total time for conducting an evaluation includes the planning process, data collection, data analysis, and presentation of the results. Data analysis and presentation are the components that make the whole process worthwhile, and sufficient time should be allotted even if this means limiting the evaluation goals and reducing the number of data collection methods.

Coding

Data collected from your evaluation must be compiled, coded, and entered into a spreadsheet or other data analysis program for analysis. *Coding* means that numbers are assigned to responses. The following example shows numbers assigned (coded) for responses to a closed-ended question:

Example:
I am able to use PubMed to avoid falling behind current medical knowledge.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Coding is typically used to analyze close-ended questions that have predetermined response categories. You can code open-ended questions, but it can be difficult and time consuming because the answers will vary with each individual response. You must read answers item by item for "naturally" occurring categories found in commonly mentioned themes. The responses are then coded according to these categories.

Quality control

Data entry must be checked for errors before proceeding. Obvious errors will be detected by scanning the entire data file (e.g. you might see a "9" when the highest possible code is a "7"). Also, ask someone who did not enter the data to compare 10% of the raw data (e.g. the surveys) with the computer data file. If there are a number of errors, all the data should be re-examined.

For the most rigorous quality control, the same data should be entered twice by different people and compared. If the compared files appear to be identical, there is greater assurance that the data were entered consistently.

Types of Analysis

The type of data analysis will vary depending on the type of data collected. Qualitative methods of data collection may include observations, interviews, focus groups, and analytic insights or interpretations that occurred during the data collection. This descriptive text is recorded and analyzed for themes. Careful reading and summarization of the data can be sufficient for general evaluation purposes (6).

There is software available for in-depth analysis of qualitative data, such as ATLAS/ti and NUD*IST. These software packages work with textual documents, such as transcripts of interviews or focus groups, and facilitate coding, search and retrieval, and theory building. NUD*IST is best known in its Macintosh version, while ATLAS/ti is most user-friendly on a DOS-based computer.

Quantitative methods of data collection use hard data (e.g. numbers of outreach participants, total Website hits) or pre-coordinated responses on questionnaires that can be coded and entered into a statistical analysis program such as SAS or SPSS.

Spreadsheet programs (e.g. Excel) can also be used to display quantitative data. Although statistical analysis is limited, it is possible to manipulate the data and produce various tables, such as frequencies, or cross tabulate the data so that relationships can be examined (e.g attitude changes in physicians vs. nurses).

Statistical techniques that summarize and describe characteristics of a group or make comparisons of characteristics between groups are *descriptive* statistics (7). If generalizations are inferred about a population based on a sample, you use *inferential* statistics.

To analyze your results, you assess the effects of your "independent variable" (the intervention) on your "dependent variables" (outcome measures). Typically, the dependent variables will be measured on your posttest survey and will include things like attitudes, intentions to act a certain way, or reports of certain behaviors.

If you were using an experimental or quasi-experimental design, the effects of an independent variable on a dependent variable would be compared between two or more groups. The independent variable (e.g. endorsement, support, and participation by opinion leaders) would only be used in the experimental group, but the dependent variable (e.g. perception of efficacy) would be assessed in both. If there are significant differences in the dependent variables between groups, you can be more confident that the independent variable made a difference.

Other dependent variables can be assessed without input from the subject. For example, you could tally how many log-ins or how much time individuals or groups spent on the computer. Then, you would determine the mean of the number of log-ins or the number of minutes spent on the computer by group. Finally, you would compare these means for significant differences, using the t-test or F-test.

T-tests

The **t-test** is a test to see if there is a statistically significant difference between the mean scores of two groups (8). For example, between an intervention group and a control group, the comparison could be the difference in mean scores on the variable "self-efficacy." To apply a t-test to the difference between the mean scores of each group, use a statistical software program such as SPSS that will use a formula to compute a t-value, or the difference between the mean scores. The program will show **t-test** results, which designate whether the t-value is larger than would be expected if the differences were due to chance. In other words, the t-test indicates whether the scores

in the intervention group were significantly different from the control group.

The t-test is particularly useful for analysis when sample sizes are small, though it is best to have at least twenty cases to compare. An **F-test** does the same thing for three or more groups.

T-tests can be used on paired samples or independent samples. In paired samples, the changes are being compared in the same individual from one point to the next (e.g. changes in attitude due to outreach participation). In independent samples, two or more separate groups are measured for comparison (e.g., outreach participants with a control group).

Univariate analysis

For some types of evaluation, descriptive data, such as background characteristics, attitudes, knowledge, and behavior, are all that is needed to describe participants. Commonly, descriptive data analyze one variable (example?) – hence the term *univariate analysis*. Descriptions are provided in terms of percentages and measures of central tendency, i.e., mean, median, and mode.

Mean – arithmetic average of all scores

Median – midpoint of all scores

Mode – the most frequently occurring score

Other examples of descriptive data are frequency or summary counts, such as the number of participants in a class.

Evaluation questions that focus on testing a hypothesis about relationships between variables require more elaborate techniques, known as *bivariate* and *multivariate* analysis.

Bivariate analysis

McKenzie (1997, p. 255) presents the following definitions of statistical techniques used in bivariate analysis.

Correlation establishes a relationship between two variables. Correlation is expressed as a value between +1 (positive correlation) and –1 (negative correlation), with 0 indicating no relationship between the variables. Correlation only indicates a relationship; this technique does not establish cause and effect.

Inferential data analysis uses statistical tests to draw tentative conclusions about the relationship between variables. Conclusions are drawn in the form of probability statements, not absolute proof. The evaluation question is stated in the form of a hypothesis. A **null hypothesis** holds that there is no observed difference between the variables (e.g., experimental and control groups' knowledge of computers). The **alternative hypothesis** says that there is a difference between the variables.

Analysis of variance (ANOVA) compares the difference in means of two or more groups. ANOVA does not prove that there is a difference between groups; it only allows you to reject or retain the null hypothesis, then make inferences about the population.

Chi square tests hypotheses about frequencies in various categories. This technique uses categories that can be distinguished from one another but are not hierarchical. Chi square could be used to analyze attitudes toward computers between physicians in three different specialties.

Multivariate analysis

Multivariate analysis determines the relationships between more than two variables. One type of multivariate statistic is **multiple regression**, used to make a prediction from several variables. For example, Gorman (1995) used multiple

regression to analyze 12 factors expected to motivate information seeking by physicians, and determined that two were significant predictors.

References:

1. McKenzie JF, Smeltzer JL. Planning, Implementing, and Evaluating Health Promotion Programs: a primer. Boston: Allyn and Bacon, 1997.
2. King JA, Morris LL, Fitz-Gibbon CT. How to Assess Program Implementation. (Second ed.) Newbury Park: Sage Publications, 1987. (Herman JL, ed. The Program Evaluation Kit; vol 3).
3. Isaac S, Michael WB. Handbook in research and evaluation : a collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences. (3rd ed.) San Diego, Ca: EdITS Publishers, 1995.
4. Reisman J. A Field Guide to Outcome-Based Program Evaluation. Seattle: Organizational Research Services, Inc., 1994.
5. Orlandi MA. Cultural competence for evaluators : a guide for alcohol and other drug abuse prevention practitioners working with ethnic/racial communities. Rockville, MD: U.S. Dept. of Health and Human Services,Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, Office for Substance Abuse Prevention, Division of Community Prevention and Training : Distributed by OSAP's National Clearinghouse for Alcohol and Drug Information, 1992.
6. Marshall JG. Using evaluation research methods to improve quality. Health Libraries Review 1995;12:159-172.
7. Hafner AW. Descriptive Statistical Techniques for Librarians. (2nd ed.) Chicago: American Library Association, 1998.
8. Fitz-Gibbon CT, Morris LL. How to Analyze Data. Newbury Park, CA: Sage Publications, 1987. (Herman JL, ed. Program evauation kit; vol 8).

Additional Resources

Dillman, D. A. (1978). Mail and telephone surveys : the total design method. New York, Wiley.

Campbell, D. T. and J. C. Stanley (1963). Experimental and quasi-experimental designs for research. Chicago, R. McNally.

Gorman, P. N. and M. Helfand (1995). "Information seeking in primary care: How physicians choose which clinical questions to pursue and which to leave unanswered." Medical Decision Making 15: 113-119.

Activities, Best Practices, Theory-based Strategies	**What will be measured?**	**How will we measure it?**

Objectives	What outcome will we measure?	How will we measure it?