

Chapter 10: Analyzing Evaluation Data

This chapter provides information relevant to analyzing the evaluation data. Discussions include the importance of selecting a data analyst, the steps in data analysis, interpreting the results, and generalizability.

The evaluation team should include a statistical analyst to help develop the data collection procedures, select the instruments for the evaluation, and conduct the analyses. The instruments selected affect the types of questions that can be asked and the types of analyses that can be performed. If no one on the team has these skills, an outside consultant should be hired to perform these duties.

Data Analysis

There are five steps to analyzing the data:

1. Cleaning the data.
2. Tabulating the data.
3. Conducting the core analyses.
4. Analyzing the data by key characteristics.
5. Interpreting the results.

Cleaning the data

Errors are likely to be made while transferring data from a questionnaire to a computer, and errors in data entry can cause faulty results. Therefore, it is necessary to “clean” the data. Data cleaning ensures that the numbers respondents indicated on the questionnaire match the numbers

entered into the computer. Data can be cleaned in one of several ways.

An individual doublechecks the entered data. One individual checks the entered data (either on the computer or on a printout of the data) against the responses on the original questionnaire. Errors are noted and then corrected.

Two individuals doublecheck the entered data. One individual reads aloud the entered data (either on the computer or on a printout of the data), while the other person checks the responses on the respondent’s original questionnaire. Errors are noted and then corrected.

Frequency analyses are conducted. A frequency analysis helps determine whether the entered numbers are within the range of possible responses on the questionnaire. For example, if the questionnaire contains a five-point Likert scale (from one to five), then the range of numbers for each question should be only between one and five. If a frequency analysis finds a question with a range of responses between one and eight, then it can be concluded that an item was entered incorrectly. However, this check ensures only that respondents stayed within the rating scale limits and not that the data were entered accurately.

A logic check is conducted. A logic check involves determining whether answers to various questions make sense. For example, if a respondent indicates that he or she has no children, all subsequent questions regarding children should have a

code of “not applicable.” Any other response suggests a data entry error.

An accepted practice is to select at random 10 percent of the questionnaires. (See <http://www.randomizer.org/form.htm> for a random numbers table.) If these contain no or very few errors, one can be relatively confident that the data are clean. However, if pervasive mistakes are found, the entire data set will need to be checked and corrected. Tracking who enters the data may identify patterns of error associated with each data entry person.

Tabulating the data

Before calculating the core analyses, the evaluator should become familiar with the data. The best way to do this is to run frequencies on the data, which give such information as how many participants completed each question and the range of responses to those questions.

Typically, data are grouped to form summaries rather than to focus on a particular individual. For example, reporting on the number of participants in an evaluation simply means counting the total number of participants who completed questionnaires. However, reporting the percentage of people who agreed to participate in the study requires dividing the number of participants who completed a questionnaire by the total number of people invited to participate in the evaluation. Percentages are often preferable to averages because, depending on the response rate, averages can be affected by a few very high or very low scores.

Scoring instruments. Several standardized or commercial instruments used in evaluations require some manipulation to create a score for each participant. For example, the Conflict Tactics Scales (Straus et al. 1996) require summing certain items to create a score for each person. When selecting an instrument for

an evaluation, be sure to obtain the instructions for scoring the instruments, regardless of who will analyze the data. Some instruments, such as the Child Behavior Checklist (Achenbach 1992), have available a computer program that scores the instrument.

Assigning weights to questions. Not all outcomes are equally important; therefore, certain questions may be weighted to have a greater effect on the results. For example, a question may have a weight of 1.5 if the outcome is particularly valuable, and a weight of 1.0 if the outcome is simply expected (Yates 1996).

Conducting the core analyses

Once the preliminary work is done, the core analyses can begin. The evaluation team should have decided during the planning stages which analyses to conduct; otherwise, once the data are collected, the great temptation is to conduct numerous analyses, which becomes unwieldy and overwhelming. The better strategy is to develop hypotheses (see chapter 5), plan the analyses around these hypotheses, and stick to the plan.

Rather than conducting the analyses only after all the data have been collected, analyses should occur periodically throughout the evaluation (e.g., monthly, quarterly). For example, first-quarter analyses can have several uses:

Enhancing adherence to the evaluation plan. Analyses conducted early in the evaluation can demonstrate that the evaluation is going to provide useful information, thus enhancing the team’s commitment to the evaluation.

Determining the need to make corrections and changes. Analyses conducted early in the evaluation can reveal whether changes in the protocol need to be made before the evaluation is complete.

Determining why discrepancies in the protocol have occurred. Periodic reports may suggest the need for reminders to individuals involved in the evaluation about why adherence to the protocol is critical, as well as possible incentives for compliance, including peer recognition and rewards.

In addition to periodic reports of the analyses, the data analyst and evaluation team should meet regularly to discuss emerging findings. These meetings could be separate from other meetings or incorporated into regular meetings (e.g., evaluation team meetings, staff meeting, multidisciplinary team meetings). Be sure to invite discussion from the team members about the results. However, keep in mind that these results are preliminary and may change with the inclusion of the entire sample. Similarly, a chance difference that appears early may disappear by the end of the evaluation. Therefore, major decisions should not be based on periodic reports (Boruch 1997).

Analyzing the data by key characteristics

If a Child Advocacy Center (CAC) has information on subgroups of individuals, for example, certain ethnic groups or children who have testified in court, the data can be analyzed by subgroup. While this may make the analyses more complex, it will also yield more realistic and meaningful results. The most useful evaluation incorporates subgroup analyses to ask the following questions:

- What works about the program?
- For whom is the program most beneficial?
- Under what conditions is the program most beneficial?

Results based on subgroup analyses will help fine tune the program. For example, differences between ethnic groups on levels of child stress during a medical examination may indicate the need to adjust the protocol to accommodate the needs of the various subgroups. On the other hand, finding no differences between these groups would suggest that the protocol is affecting all clients equally.

Interpreting the results

Interpreting the results is often the most difficult aspect of any evaluation for several reasons, discussed below.

Numerical context and explanation.

Numbers typically need to be placed in some context for their meaning to be discernable. Consider the following example:

CAC Alpha shows an increase in prosecution rates from 35 percent to 50 percent, which is pretty good.

CAC Beta shows an increase in prosecution rates from 5 percent to 20 percent, which is great.

Both examples show an increase of 15 percent in prosecution rates, and yet it is very different to be starting at 35 percent instead of 5 percent. The reader needs a context within which to interpret the numbers.

As another example, what does it mean to say that a center has served 300 children this year? Whether this is a lot or a little depends on the context in which the center operates. If CAC Alpha reported that there were 5,000 reports of child sexual abuse (CSA) in the counties that it serves, and the center served 300 of those children, the reader knows that the center is serving a small percentage (6 percent) of the children who allege that abuse has occurred. In contrast, if there

were 500 CSA cases in the counties that CAC Beta serves, and the center served 300 of those children, the reader knows that it is serving a large percentage (60 percent) of the children who allege that abuse has occurred. Thus, a CAC could be serving a few or a lot of children, but there is no way to know which without a numerical context. Numbers in isolation are basically meaningless.

“One of our outcomes was to increase the number of families who actually go into therapy. We found that 75 percent of families say they want to enter therapy, but only 30 percent actually do. Why don’t they? What’s going on here? This suggests some missing link here. Now we have to find the missing link.”

Not only do numbers need a context, they also require explanation to help readers understand what they mean. An explanation answers the question why—what accounts for these results? For example, the finding that 6 percent of the CSA cases are referred to a center can be explained in two ways. It could mean that the center is not serving very many children. However, another interpretation is that most of these cases are not being referred to the center. The question, then, is why not? With this information, agencies can then determine why agencies are referring so few cases to the center.

If the evaluation results differ from the predictions, this discrepancy must be explained. When thinking about possible explanations, always consider internal and external influences on the evaluation. For example, possible external influences on the results may include rising unemployment in the neighborhood or reduced funding for the program. Possible internal influences may be high staff turnover or the introduction of a new curriculum.

Implications and recommendations.

Another difficult evaluation task is to derive implications from the findings: What can be inferred from these findings? It is insufficient to simply state a conclusion (i.e., a statement or a set of statements about the merit, worth, or value of the evaluation) without addressing the implications of that conclusion. For example, what are the implications of finding a drop in referrals for a particular ethnic group? Management might want to replace the director of program services, but the evaluator might want to conduct a followup study to determine why the drop in referrals occurred. Be sure to discuss with the team members the possible implications of the findings.

The team should discuss the implications of the findings because recommendations flow most naturally from the implications. Some of the exercises discussed at the end of chapter 3 can facilitate these discussions. The team will need to make explicit recommendations for the evaluation report because more often than not, data do not speak for themselves. In addition, even if the readers could form their own recommendations, they should also receive the evaluation team’s recommendations, as the two sets of recommendations may differ. However, it is a considerable leap from conclusions to recommendations, so be cautious in making recommendations (Scriven 1993).

Statistical significance versus practical significance.

Statistical significance refers to whether results occurred at a level greater than chance. Some events occur due to chance alone; therefore, a test is needed to determine whether the results were due to chance or whether the probability of a particular result occurred at greater-than-chance levels. Researchers have long agreed that there is statistical significance if the probability of the result occurring from chance alone is less than 5 percent (denoted by $p < .05$).

One shortcoming of relying on a significance level is that it depends on the number of participants in the evaluation. That is, it is far easier to reach significance with a large number of participants (i.e., a large sample size). Therefore, some researchers have started to report critical intervals rather than significance levels. Critical intervals indicate the degree of confidence one can have in the results when they fall within a particular range. In one example, there is a correlation of .63 between case review and the case being accepted for prosecution, and the confidence interval is 95 percent. One can be 95 percent confident that the result (the correlation) is not due to chance if the correlation falls between .61 and .65. That is, in 95 out of 100 samples from the same population, the estimated correlation should fall between .61 and .65.

Although researchers adhere to statistical significance, statistical significance and practical significance may be different. That is, statistical significance does not always reveal the importance of the result. For example, differences that are very small are not likely to be important, even if they are statistically significant (remember that significance is strongly affected by the number of participants in the evaluation). As a rule of thumb, differences of less than 5 percentage points are seldom meaningful for program managers or funding agencies. Differences of 10 or more percentage points are more likely to be of practical concern (United Way of America 1996).

Finding no differences. Directors are often concerned that an evaluation will fail to reveal the program's effectiveness. However, lack of significant change among the participants, for example, does not necessarily rule out program effectiveness (Boruch 1997). Below are several possible explanations of why an

evaluation failed to reveal program effectiveness:

- *Differences may exist, but the data do not reflect this fact.* Often the program works differently for different people, and analyzing data only for the group of participants as a whole may not reveal differences. One way to test for this is to include in the analyses a measure of something that could affect the results (referred to as a *moderating variable*; see chapter 6). For example, if child age is a potential moderating variable in the analysis of child stress, older children may demonstrate significant differences in pre-post intervention levels of stress, while younger children may not.
- *The measurement of the response to the program was invalid.* Often instruments are blamed when no differences are found, particularly if the measure was developed by the investigator for a particular study, and therefore the validity and reliability are unknown. It may be that the instrument does not measure what the team intended to measure (in technical terms, the instrument is not valid). For example, a child behavior scale would not be a valid measure of child stress because it measures child behavior and not child stress.
- *The statistical power of the experiment is too low.* Statistical power refers to the probability of detecting differences in the effectiveness of the program. Fewer than 7 out of 10 studies are sufficiently powerful to detect differences of even moderate size. "No difference" results are a real possibility. However, one can ensure having enough statistical power to detect differences by conducting a power analysis (Cohen 1992a). In addition, recruiting participants who are similar on some important characteristics (referred to as "homogeneity")—for example, by

recruiting participants who are all victims of CSA—reduces the amount of variability among participants and therefore increases statistical power.

- *The wrong population participated in the evaluation.* This is less likely to occur at a CAC. However, data analysis may reveal no differences if, for example, the dysfunctional families are excluded from the study because they refuse to participate, they drop out of the program, or staff are unable to locate them at a later date, leaving only more functional families participating in your evaluation. Functional families may not benefit from the CAC's services as much as dysfunctional families, and therefore the evaluation would not find significant changes among functional families.

A number of factors may explain a finding of no difference, and sometimes the results will not be as expected.

Typically, several factors may explain the evaluation's results. Therefore, select a theory (or process) for why certain results may occur before implementing the evaluation and eliminate as many competing explanations as possible by measuring competing explanations (see chapter 5). For this reason, the evaluation should include the following:

- *Exposure to other important influences.* Chapter 8 discusses a number of contexts to consider when planning an evaluation. This might help determine which contexts could influence the results.
- *Program monitoring evaluation.* To ensure that the outcomes result from the program rather than from some other factor that was not measured, simultaneously conduct a program monitoring evaluation to ensure the services that were supposed to be provided to clients actually were provided.

Recruitment challenges: Voluntary participation and attrition. Voluntary participation refers to a sample selection method in which participants in the evaluation consist only of those individuals who voluntarily agree to participate. Many directors conducting client satisfaction surveys, for example, report difficulty obtaining information from every client and, therefore, data collection is limited to those individuals willing to participate. Although not purposefully selecting success-prone participants for the evaluation (known as “creaming”), by having data only on these voluntary participants, the program may appear more effective than it really is. Participant attrition, on the other hand, refers to individuals who started the program (and therefore some data may have been collected on them), but who fail to complete the program or are unable to be contacted later for followup data collection. As with voluntary participation, an evaluation report based on data collected only on individuals who completed the program or who were available for followup data collection may make the program appear more effective than it really is. More important, implications and recommendations based on information received from this limited pool of clients may be misleading, and even damaging, to the program.

Generalizability. Typically, a researcher selects a subset of individuals (referred to as the sample) from a total pool of individuals (referred to as the population) to participate in a study or evaluation. For example, a center might randomly select 25 percent of the clients seen at the CAC to complete a client satisfaction survey rather than requiring 100 percent of the clients to participate. The assumption is that the results from this random sample generalize to the population (that is, the sample is representative of the population of CAC clients). The results of an evaluation based on a representative subset of

participants would be the same if the evaluation included all CAC clients. Whether a study's results are generalizable depends heavily on the sample selection method and what questions are being asked.

For example, to learn how law enforcement personnel on the multidisciplinary team perceive the CAC, one should ask a subset of those law enforcement personnel who interact with the CAC to participate in the evaluation. To learn how law enforcement in the larger community perceive the CAC, one should ask a subset of all law enforcement in a particular jurisdiction to participate in the evaluation. These are very different samples of law enforcement that are perfectly appropriate for each of the questions being asked.

As another example, whether 10 percent of all reported CSA cases referred to a CAC is generalizable to all CAC cases depends on whether the 10 percent of cases referred to the CAC were similar to all CSA cases reported in the jurisdiction (making the results generalizable), or whether that 10 percent of cases represented only the most egregious CSA cases (making the results not generalizable).

Generalizability is hampered by a voluntary participation recruitment strategy because those who decline to participate in an evaluation may be systematically different from those who agree to participate (e.g., more serious cases, greater family dysfunction). An effect based on the voluntary sample may indeed hold for people like those in the voluntary group, but it cannot be determined whether the effect holds for the entire client population. Thus, defining eligibility criteria of potential participants is essential for understanding the generalizability of the evaluation (Boruch 1997).

One strategy for assessing the effect of attrition and voluntary participation on the evaluation results uses the data collected (e.g., on intake forms) from individuals who refuse to participate, who drop out, or who cannot be contacted for the followup to identify any differences between those individuals and individuals who agreed to participate in the evaluation. If differences are found, it may be argued that the program would be deemed less effective if all CAC clients were included in the evaluation. On the other hand, if no differences are found between the two groups, then there can be greater confidence that the evaluation results are generalizable.